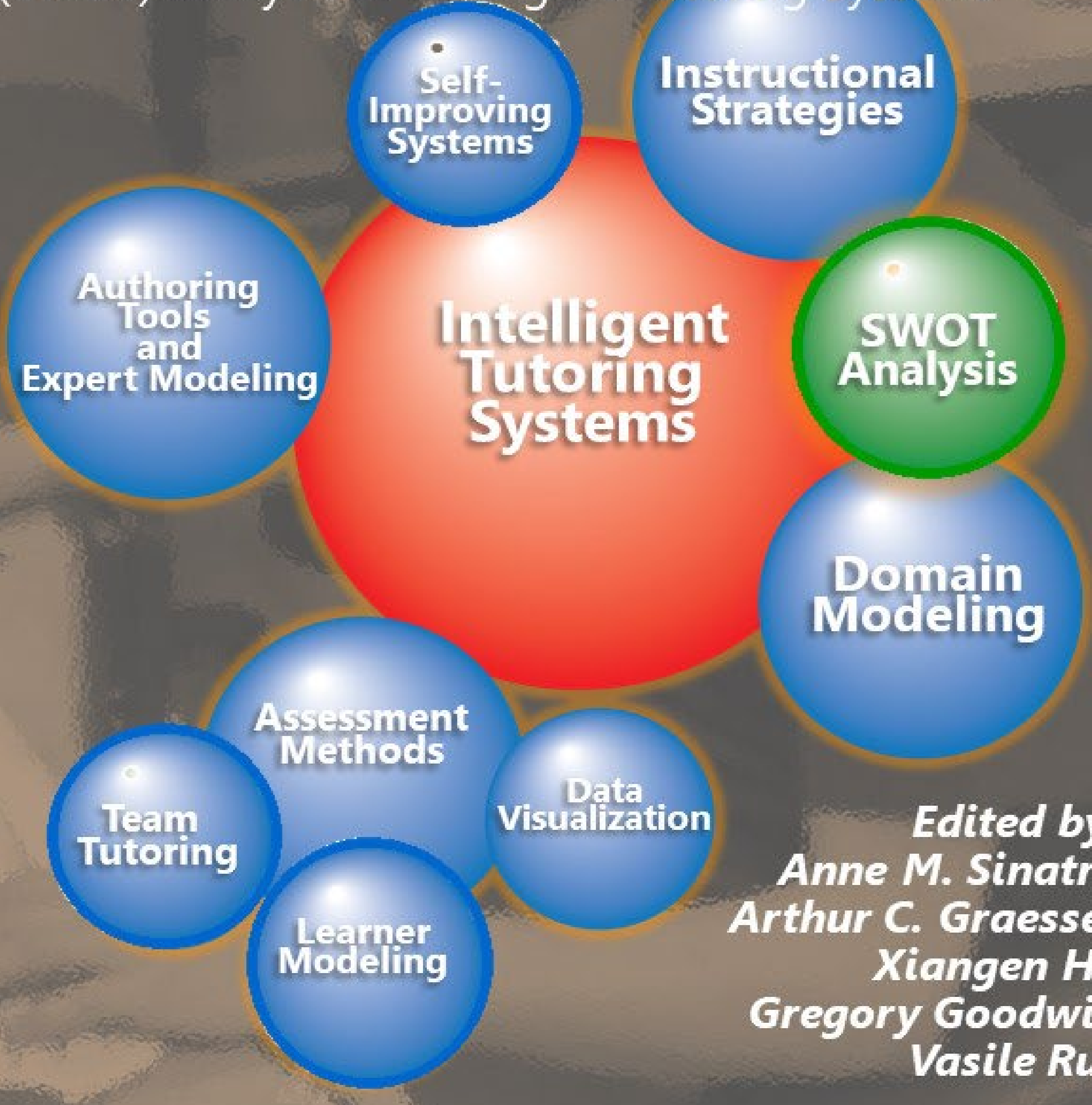
Design Recommendations for Intelligent Tutoring Systems

Volume 10
Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems



Edited by:
Anne M. Sinatra
Arthur C. Graesser
Xiangen Hu
Gregory Goodwin
Vasile Rus

A Book in the Adaptive Tutoring Series

Design Recommendations for Intelligent Tutoring Systems

Volume 10
Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems

Edited by:
Anne M. Sinatra
Arthur C. Graesser
Xiangen Hu
Gregory Goodwin
Vasile Rus

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, March 2023

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

International Standard Book Number:
978-0-9977258-3-4

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 SWOT ANALYSES

Anne M. Sinatra[1], Arthur C. Graesser[2], Xiangen Hu[2],
Gregory Goodwin[1], and Vasile Rus[2] Eds.

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

This book is a resource for those who are new to intelligent tutoring systems (ITSs), as well as those with a great deal of experience with them. This is the tenth book in our ***Design Recommendations for Intelligent Tutoring Systems*** book series. The focus of this book is on Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analyses of varying components of ITSs. Each chapter in the book represents a different topic area, and includes a SWOT analysis that is specific to that topic and how it relates to ITSs. This book can be read in order, or a reader can choose a specific topic area and move directly to that chapter.

Each SWOT Analysis describes the current state of the topic area, and how the lessons learned from the analysis could be applied to the Generalized Intelligent Framework for Tutoring (GIFT) (Sottilare et al., 2012; Sottilare et al., 2017). GIFT is an ITS architecture that is open-source, modular, and domain independent (Sottilare et al., 2017). Each book in the design recommendations series has addressed a different ITS topic area, and how the work in each chapter can relate to and inform the GIFT architecture. GIFT has continually been in development, with features consistently being added to improve functionality, as well as reduce the skill requirement for authoring content in GIFT. GIFT is freely available in both downloadable and Cloud versions at https://www.GIFTtutoring.org.

There have been a series of yearly Expert Workshops that started in 2012 as part of a cooperative agreement between the University of Memphis and US Army Research Laboratory (in 2018, the GIFT team as part of a reorganization, became part of US Army Combat Capabilities Development Command – Soldier Center). These workshops each had a relevant ITS topic area, and included invited experts from academia, government, and industry. Each workshop led to a book in the ***Design Recommendations for Intelligent Tutoring Systems*** book series. These books captured the themes of the workshops in the form of collaborative chapters between experts who participated.

The tenth expert workshop topic was SWOT Analyses of Intelligent Tutoring Systems. This workshop was structured in line with the topic areas that the first 9 expert workshops covered (see Table 1). There were presentations on each of these areas and in most cases two experts each presented their SWOT analysis of the topic area. Additionally, there was an overview presentation on GIFT, and an overview ITS SWOT analysis. Table 1 indicates the topics of each workshop, and the associated book publication.

Table 1. List of the topics of the first 9 expert workshops, the workshop date, and the book publication date.

Topic	Workshop Date	Book Publication Date
Learner Modeling	September 2012	Volume 1 – July 2013
Instructional Management	July 2013	Volume 2 – July 2014
Authoring Tools	June 2014	Volume 3 – June 2015
Domain Modeling	June 2015	Volume 4 – July 2016
Assessment Methods	May 2016	Volume 5 – June 2017
Team Tutoring	May 2017	Volume 6 – August 2018
Self-Improving Systems	May 2018	Volume 7 – October 2019
Data Visualization	August 2019	Volume 8 – December 2020
Competency-Based Scenario Design	September 2020	Volume 9 – February 2022

Sections of the Book

This book is organized into three sections that cover SWOT analyses in different groupings:

I. GIFT and Intelligent Tutoring Systems

II. Intelligent Tutoring System Components

III. Advanced Elements of Intelligent Tutoring Systems

Section I covers a general overview including a GIFT SWOT analysis (including the history of GIFT), and a general ITS SWOT analysis. Section II is made up of SWOT analyses of traditional components of ITSs and includes: Learner Modeling, Instructional Strategies, Authoring Tools, and Domain Modeling. Section III covers SWOT analyses of advanced elements of ITSs including Assessment Methods, Team Tutoring, Self-Improving Systems, Data Visualization, and Competency-Based Scenario Design.

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 – GIFT and Intelligent Tutoring System SWOT Analyses

SWOT Analyses of:

Generalized Intelligent Framework for Tutoring (GIFT)

Intelligent Tutoring Systems

Chapter 1 – Generalized Intelligent Framework for Tutoring (GIFT) SWOT Analysis

Benjamin Goldberg and Anne M. Sinatra
U.S. Army Combat Capabilities Development Command (DEVCOM) Soldier Center

Introduction

The focus of the 2021 Soldier Center/University of Memphis expert workshop was centered around a SWOT (Strengths, Weaknesses, Opportunities, and Threats) Analysis of Intelligent Tutoring Systems (ITSs) and the functions that make those platforms work. During the three-day event experts in the field presented their SWOT perspectives across numerous research and development themes. Each topic area was tied to a previous Design Recommendations for Intelligent Tutoring Systems book volume theme, including: Intelligent Tutoring Systems in general, Learner Modeling, Instructional Management, Authoring Tools, Domain Modeling, Assessment, Team Tutoring, Self-Improving Systems, Data Visualization and Competency-Based Scenario Design.

Across the better part of the last decade, this aforementioned series of books were used to assist in guiding the design and implementation of the Generalized Intelligent Framework for Tutoring (GIFT). Ultimately, GIFT was designed as a domain-agnostic and environment-independent architecture based on documented best practices. To drive its implementation, a set of standard tools, data models, and workflows were established to guide the Adaptive Instructional System (AIS) creation process. All of the technologies have evolved over the years based on the use cases and stakeholders driving their development, and we have made significant progress since the initial public release of GIFT back in 2011.

But how far have we come, and what challenges lay ahead? In this chapter, we do our own critical SWOT Analysis of the GIFT architecture and reflect on the successes and challenges experienced across the execution of this program. In GIFT, instead of needing to recreate the entire AIS infrastructure for each implementation, the framework stays constant, and the content can be changed. There are authoring tools in GIFT that are used to add and create content as part of the ITS. Each of the 9 previous workshops and book volumes not only discussed the general research area, but also provided recommendations for how to improve GIFT in context of the topic. Similarly, the 2021 workshop included discussions of how the SWOT Analyses specifically can be used to improve GIFT. In this chapter, we go an additional step and conduct a SWOT Analysis of GIFT itself.

Background and Supporting Research

The evolution of GIFT has been heavily influenced by the use cases and learner populations engaged across the program's history. A big emphasis of GIFT's development was to establish a set of generalizable tools and best practices that harness the benefit of intelligent tutoring and apply them explicitly across military relevant skill and competency domains. These technology objectives were documented within the Army Learning Model 2015 document (Army, 2011), and directly justified a research investment to build capabilities to meet future Army training and education requirements.

To support these objectives, specific research vectors were established that influenced the set of capabilities examined and iteratively developed, with established use cases creating the context to guide project execution. These vectors looked at elements across the core components of an AIS, such as: (1) how to model the learner and team within a learning and military organization, including elements of their cognitive, physical, and emotional ability, (2) how to model the domain and task environment used to infer performance and track proficiency, (3) how to model pedagogy (i.e., the art of instruction) and establish coaching agents that can manage feedback and adapt the experience, and (4) how to manage all of these interconnected processes across an ecosystem of learning and simulation resources that span Live, Virtual and Constructive (LVC) type interactions and data sources. Figure 1 provides a timeline of GIFT use cases and development, which are described in more detail throughout this chapter.

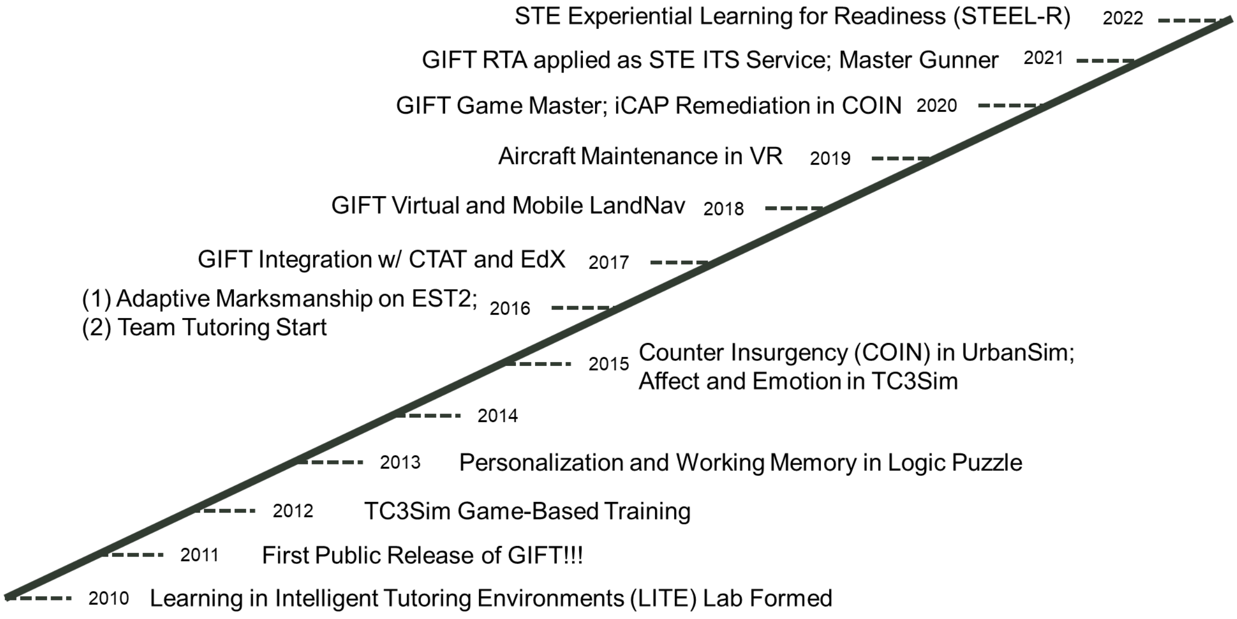


Figure 1. Timeline of GIFT Use Cases and Development.

Learning in Intelligent Tutoring Environments Lab (LITE) Formed

The adaptive training research program kicked off in 2010 through the creation of the Learning in Intelligent Tutoring Environments (LITE) Lab at the Simulation and Training Technology Center (STTC). It started as a small team of four, with a near-term focus on establishing a modular and extensible architecture that would be applied against all research questions investigated by the LITE Lab (Sottilare et al., 2012). As a guiding use case, the first GIFT proof of concept looked at supporting automated assessment and adaptive training in the Army's legacy Games for Training program.

First Gateway Module to Interoperate GIFT with an External Training Environment

The first gateway module to interoperate GIFT with an external training environment was created for Virtual Battle Space 2 (VBS2). A scenario was designed with the VBS2 mission editor and involved a trainee executing a patrol around an identified compound. The scenario provided an excellent sandbox to experiment with different data driven techniques to monitor interaction within a simulation environment, and to assess that interaction against real-time performance criteria. In this instance, we designed and implemented GIFT condition classes (i.e., re-usable conditional logic configured with scenario specific parameters) based on game state information extracted from Distributed Interactive Simulation (DIS) standardized data packets, with the use of the VBS scripting language. This allowed us to establish the message sets across all core GIFT modules to support capture of data, use of data to assess performance, and performance assessments that drive real-time feedback and adaptation strategies (e.g., change the weather from clear to foggy).

To further improve on this VBS integration, another VBS scenario was applied using an Army Games for Training program of record validated Training Service Package (TSP). This scenario involved clearing a building of enemy combatants and added additional condition classes to handle engagement tasks that measured and assessed tactics and behaviors. The resulting exemplar further demonstrated that the approach being applied in GIFT was easily extensible across scenarios and task domains supported in a single training environment. These scenarios were used to iteratively refine the data communicated across GIFT's modules, and to better understanding the logic required to translate game state information into valid metrics used to monitor and assess performance. For more information on GIFT's modular design and implementation during the earlier years in the program, see Sottilare et al. (2012).

Integration with Tactical Combat Casualty Care Simulation (TC3Sim) Serious Game

With a baseline infrastructure in place that supported a single training application, it was time to expand to a new use case to support the domain-independent and environment-agnostic requirements linked to our program, as well as to drive forward on our first empirical evaluation examining the influence of AIS technologies in a military skill domain. In 2012, we selected a second Games for Training product to develop within called TC3Sim (TC3 stands Tactical Combat Casualty Care). The game was developed to provide first-person type exposure to scenarios and events that require the knowledge and procedural application of combat lifesaving skills (e.g., triage, hemorrhage control, burn care, preparation for transport and medevac, etc.). A socket-connection was created with TC3Sim through a Simple Object Access Protocol (SOAP) Gateway specification, enabling real-time capture and routing of game play data through GIFT's core modules. We then designed 23 automated assessments linked to a mission context that challenged two types of medical assistance: Care Under Fire and Tactical Field Care. The authored assessments were validated with Subject Matter Experts and then aligned to coaching prompts within a GIFT Domain Knowledge File (DKF). For this development, we highlighted GIFT's first integration with an external assessment engine (i.e., a data-driven service developed outside of GIFT). The DKF worked with a tool called SIMILE to manage and control the real-time performance classification (Mall & Goldberg, 2014). We collaborated closely with the United States Military Academy (USMA) at West Point, and conducted an empirical evaluation examining the impact of real-time assessment and coaching on transfer performance. We also further investigated whether the modality of coaching had a significant effect on the performance variables we were monitoring. Results highlighted a significant improvement in performance when receiving real-time coaching via GIFT. The results also showed benefit for inclusion of a pedagogical agent to serve as the feedback delivery vessel, as that method for system interaction follows principles and heuristics informed by Social Cognitive Theory (Bandura, 2001). For a full review, see Goldberg and Cannon-Bowers (2015).

Personalization Research and the Logic Grid Puzzle Tutor

In the same 2013 timeframe, an additional experiment was conducted using GIFT to examine the impact of personalizing materials during the tutoring process (Sinatra et al., 2014). A logic grid puzzle tutor was created using Visual Basic for PowerPoint and macros. A PowerPoint Show with macros was included as part of a GIFT course that taught participants how to solve logic grid puzzles, and then asked them to answer questions and solve additional puzzles afterwards. There were three different versions of the PowerPoint tutor that were created which represented different conditions: self-reference, popular media, and generic. The logic puzzle tutor included different names in it during the learning phase depending on the condition. In the self-reference condition the participant entered their own name and the names of friends; in the popular media condition character names from the Harry Potter series were included; and in the generic condition, general names were included. Consistent with the self-reference effect (Symons & Johnson, 1997), it was anticipated that if the participant had a personal tie to the names included in the material (self-reference) they would perform better than if they received generic names. Further, it was investigated whether names from popular media would have a similar impact. It was found that general enjoyment of thinking/learning as represented by the need for cognition (NFC; Cacioppo et al., 1984) interacted with transfer performance (score on a difficult logic grid puzzle), such that those with a low NFC score appeared to actually be negatively impacted by the inclusion of self-referential names, which may be consistent with the seductive details effect (the inclusion of extraneous details of interest which may lead to distraction) (Harp & Mayer, 1997). There was no negative impact found for those who were high NFC. For transfer performance, there were no significant differences found for either those who were Low NFC or High NFC between the popular culture and generic conditions (Sinatra et al., 2014). The generic condition version was released with GIFT as a showcase course in GIFT that can be run as an example.

Affect and Emotion in TC3Sim

Starting in 2014, we began research examining the affective modeling requirements in GIFT and the use of sensors to drive learner model updates. With an in-place TC3Sim integration, a new project was initiated that examined both sensor-based and software-based affect detectors. A goal of the learner modeling research vector in GIFT is to utilize low-cost unobtrusive methods to monitor affective states that impact learning and retention. For this experiment, we integrated wearable physiology sensors and a Microsoft Kinect within GIFT's sensor module. To further support affective modeling, we integrated the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) to align human observed affect labels synchronized with all data sources via GIFT (Baker, Ocumpaugh & Andres, 2018). This established a multi-modal infrastructure to drive model and classifier development. New scenarios were authored in TC3Sim with scenario characteristics for inducing affective responses (e.g., adding fog to induce fear, adding incurable patient to induce frustration, etc.). Multiple data collections were executed to drive model creation, training, and validation. The states of frustration and boredom showed the highest accuracy in being detected within a trained affect classifier when evaluated through a cross-fold validation. We then integrated GIFT with an open source version of a tool called RapidMiner to drive real-time classification based on our offline trained models. With this real-time affect monitoring capability, we established pedagogical interventions that would be automatically enacted if an upward trend of frustration is detected. For a breakdown on those experimental findings, see DeFalco et al. (2018).

Counter Insurgency in UrbanSim

In parallel to the affect research described above, a new use case was established using a discrete game-based simulation developed at the Institute for Creative Technologies (ICT) called UrbanSim (McAlinden et al., 2008). The game environment was designed to train mission command type contextual scenarios that were launched within a counter insurgency campaign. We integrated the platform with GIFT with specific attention to studying self-regulated learning and practice behaviors that align with metacognitive skills and abilities in a discrete-event simulation environment (Maidstone, 2012). The goal was to study how trainees used the resources and components within the environment while executing tasks, and then to model optimal and inefficient self-regulated interaction patterns and behaviors based on performance outcomes. These metacognitive skills were modeled in a way to highlight transferability across tasks and environments, so as to train specific skills that will benefit self-regulated learning requirements. If sub-optimal behaviors are recognized, automated coaching strategies can be injected that aim to instill proper self-regulated strategies rather than provide corrective feedback. The outcome of this effort established a new hierarchical learner model to be operationalized in GIFT, with current efforts using its formalization. See Biswas et al. (2019) for an overview of the work.

Adaptive Marksmanship in the EST2

Moving forward from TC3Sim and Urbansim, two new challenging endeavors were initiated: (1) intelligent tutoring for a psychomotor skill and (2) intelligent tutoring in a dynamic team context. In the psychomotor domain use case, we leveraged the Army simulation-based Engagement Skills Trainer 2 (EST2) program of record used for initial hands-on training of Basic Rifle Marksmanship (BRM) tasks and fundamentals. A trainee interacts with a fit/form/function simulated rifle and engages targets on a projection screen through calibrated infrared sensing and hit-detection technology. The environment is a rich data source as the weapons are outfitted with embedded behavioral sensors (e.g., aim point, trigger pressure, cant angle, buttstock pressure, etc.). When integrated with GIFT, we added additional wearable sensors to examine heartrate and breathing patterns in addition to the system data types. We took this extended system to Fort Benning and worked with the world class Service Rifle Team within the Army Marksmanship Unit. Through this data source we successfully established validated models of expert performance in relation to the BRM fundamentals of steady position, proper breathing, trigger control, and aiming. We then utilized these models to inform new GIFT condition classes that analyze real-time data captured during a training run against the representative expert models (i.e., answering the question on whether or not a trainee was exhibiting proper fundamentals). This would drive adaptive coaching decisions after each training trial, enabling a personalized BRM experience. The methodology and outcomes of that process can be accessed through Goldberg et al. (2018). An empirical test on training effectiveness was conducted, but results have not yet been publicly published.

Team Tutoring

In this same timeframe, we also initiated a significant investment in the team tutoring space, which we are still working today. This involved an extensive meta-analysis/review of the literature to guide initial designs (see Sottilare et al., 2018), along with prototyping efforts to establish architecture requirements to support this new learning audience. The prototyping started with a conceptual exploration of how to represent the team formation and assessment requirements within GIFT's software baseline at the time. A "simple" two-person scenario was developed in the VBS3 game environment that challenged the distributed team to monitor and communicate (i.e., inform vs. acknowledge) the activity of non-player characters in their environment. If a hostile character (e.g., carrying a weapon) was leaving a team member's zone and entering another team member's zone, they were instructed to communicate that transfer, with an acknowledgment on the other end. This use case challenged the utility of GIFT's DKF, and led to multiple modifications and extensions to drive these interdependent assessments running in parallel. For an overview of the research and early pain points in implementation, see Gilbert et al. (2018) and Ostrander et al. (2020). This work was later scaled up to a three-player version which introduced additional challenges in assessment, and different responsibilities for some of the teammates (Ouverson et al., 2021). The team-based intelligent tutoring research continues today, with significant advancements in GIFT's domain modeling techniques, automated assessments, and team coaching functions. Work in the area of team tutoring is ongoing, and the proceedings of the tenth annual GIFT Users Symposium documents the most up-to-date implementations and works-in-progress at the time of writing of this chapter (Sinatra, 2022); the proceedings can be accessed for free at https://gifttutoring.org/documents/159.

Integration with CTAT, EdX, and MOOCs

Another area of research investment focused on the role of GIFT as an AIS resource within an enterprise level Learning Management System and using an ecosystem approach to drive the learning progression. There were two initial efforts in this area. The first examined the utility of GIFT in supporting Massive Open Online Course (MOOC) platforms (e.g., EdX, Coursera, etc.) through the Learning Technology Interoperability (LTI) Standard (IMS, 2012). In this use case, we integrated directly with Carnegie Mellon's

Cognitive Tutor Authoring Tool (CTAT) and established that system as an available course object in GIFT by making GIFT an LTI consumer. We then integrated directly with the LMS edX to enable a GIFT course as an available activity within their platform by making GIFT an LTI provider. This allows an authored GIFT course to be utilized as a resident resource to LTI compliant LMSs, while also allowing a GIFT course to utilize other LTI providers within its course structure. This approach enables a shared lesson experience that can navigate across several problem sets and scenarios, each utilizing disparate systems and technologies. Through these mechanisms, an evaluation was executed using the Big Data in Education MOOC (https://www.edx.org/course/big-data-and-education), facilitated by Professor Ryan Baker at University of Pennsylvania. For a full breakdown, read Aleven et al. (2018).

GIFT's Engine for Management of Adaptive Pedagogy (EMAP) and Data-Driven Tutorial Planning

Another focus area within the GIFT program centered on personalized course sequencing. Specifically, this focuses on operationalizing instructional design theory with underlying data analytics to assist non-technical audiences in building adaptive lesson materials at the macro-adaptation level (i.e., selecting what happens next to support a defined learning objective). This led to the establishment of GIFT's Engine for Management of Adaptive Pedagogy (EMAP). The EMAP operationalized Merrill's (2002) Component Display Theory and established configurable quadrants of learning based around a Rules, Example, Recall and Practice design paradigm. This led to a pedagogical agent that would make informed decisions on persistent data captured in a learner model, and would apply metadata to align learner attributes with evidence-based instructional strategies grounded in literature (Goldberg, Tarr, Billings, Malone, Brawner & Sottilare, 2012).

A second project in this area centered on establishing data-driven tutorial actions that take advantage of reinforcement learning techniques and create self-optimizing pedagogical features that improve over time. This project built on top of GIFT's EMAP (Goldberg, Hoffman & Tarr, 2015) by establishing a remediation phase of learning interaction. The remediation phase was delivered after an assessment, and aligned with the concepts and learning objectives that performed the weakest, thus requiring mediation. To personalize the remediation experience further, the Interactive, Constructive, Active, Passive (ICAP) learning activity framework (Chi, 2009) was operationalized across a set of Markov Decision Processes that were designed to identify what to focus the remediation on, and what type of interaction is required to manage the issue/impasse being experienced at the learning level. These techniques were implemented using multimedia learning content centered on foundations of Counter Insurgency tactics. A study was executed using Amazon's Mechanical Turk platform, with outcomes showing necessary policy shifts in remediation logic as a learner progresses through an extended time window of interaction. For a review of the work, see Spain et al. (2021).

Land Navigation and GIFT Mobile App

An additional exciting area of research we kicked off in this timeframe was examining the utility of GIFT in more dynamic and eXtended Reality (XR) environments. The initial endeavor in this area involved developing GIFT's first mobile application, which enabled experiential active learning linked to physical locations within a defined area. For this capability, we partnered with USMA and their Simulation Center and Department of Military Instruction to implement and pilot a GIFT mobile app lesson that targeted land navigation skills and procedures. The goal was to a create a personalized and self-guided "terrain walk" with embedded tasks, assessments, and adaptive coaching. As a trainee walked a defined route, the phone would vibrate and initiate specific tasks based on their location. This would require the trainee to engage in comprehensive land navigation tasks, with inputs and assessments managed through GIFT's tutor user interface. The application was evaluated within the USMA Beast Course, with 130+ New Cadets engaging with the technology instead of the traditional one-instructor-to-many-students guided terrain walk. See Goldberg et al. (2018) for an overview of the technical implementation to drive this training interaction. One highlight of the assessment outcomes was that all students qualified when formally assessed three days after the training intervention.

Aircraft Maintenance in Virtual Reality

The other endeavor in the XR space explored immersive training practices that leverage the latest in Virtual Reality (VR) head-worn devices. This effort was executed in collaboration with the Boeing Company through an established Cooperative Research and Development Agreement (CRADA). The project served two primary functions. First, we examined the extension of GIFT's assessment and pedagogical supports by establishing interoperability with Boeing's intelligent tutoring system platform. This embedded Boeing's in-house tutoring resources as an integrated course object within GIFT, enabling an additional tutoring service to interoperate within the GIFT infrastructure. This is significant, as it enables external AIS services and tools to be leveraged within GIFT, establishing an open system architecture that drives platform interoperability. The second objective was to then examine assessment supports leveraging data generated during a full VR engagement (i.e., extending game-based assessment techniques to support data and interaction types enabled through a full head-mounted virtual engagement). We leveraged the HTC Vive headset with content in Unity. This new interaction mode was applied to a use case on procedural training of performing maintenance activities on a P8 aircraft. The resulting system was used in a USMA capstone effort, with an effectiveness analysis looking at the impact of blended AIS techniques across GIFT and Boeing's immersive VR interactions. See Rea, Rengel, Buck, Goldberg and Rovira (2019) for a full overview of the training effectiveness results.

GIFT's Alignment with the Army's Synthetic Training Environment

The GIFT program was tasked to support the Army Future Command's investment in the Synthetic Training Environment (STE). The STE is being designed to update and modernize the core tools, methods, and environments the Army utilizes to deliver its collective simulation-based training requirements. This includes utilizing advancements in gaming and extended reality environment technology, while also applying data-driven functions and utilizing intelligent tutoring type services. In this instance, the initial goal was to interface a human instructor/observer to the platform, and to use different strategy and visualization techniques to help them best control the training experience.

As a starting point, this required a capability to interface a user with an AIS and provide human-on-the-loop functionality. The GIFT Game Master prototype was developed in 2019 which included a human observer controller to be involved in the GIFT interactions in real-time. The Game Master provides an interface to visualize numerous data streams being provided to and produced by GIFT while learners are engaged in an external training application scenario (e.g., VBS4). An observer can view map-based information such as unit locations and engagements, and real-time automated assessments and pedagogical decisions managed by GIFT DKF. Furthermore, the observer can actively participate by providing assessments of the learner and applying scenario injects as needed. The Game Master includes playback functionality that displays a timeline after the training scenario has been completed, which can assist with After Action Reviews. For more details on the Game Master and its integration into GIFT, see Goldberg, Hoffman and Graesser (2020).

Also in 2021, a new configuration of GIFT called Real-Time Assessment (RTA) was developed, with an emphasis on de-coupling GIFT's real-time tutoring processes from the rest of the framework. This enables GIFT to operate a data service within any open system architecture, with user experience and interface facilities managed through a separate client. This capability was delivered to STE to support upfront intelligent tutoring and adaptive training requirements native to GIFT, while removing the other support functions not associated with real-time tutoring processes. Unlike previous configurations of GIFT which require both instructors and learners to use the various GIFT webpages to manage and use GIFT, the RTA configuration deploys GIFT as a service with no user interface. The functionality GIFT provides is accessed through a socket connection managed in the GIFT Gateway module. That allows external systems to initialize GIFT with a specific DKF, then receive GIFT learner state and pedagogical request updates in real-time. These external systems can thereby decide how to handle this new stream of information and display it in their own user managed devices.

Master Gunner Course Pilot Study

During this engineering focus, GIFT was also explored on its utility within an Army institutional use case. In 2021 GIFT was used in a pilot study for the Maneuver Center of Excellence's (MCOE) accredited Master Gunner course at Ft. Benning. Instructional materials for three topic areas were utilized to create full GIFT-adaptive tutoring lessons based on GIFT's EMAP adaptive course object with ICAP remediation. The authored courses were provided to students prior to their first day of class via GIFT Cloud. This allowed a real-world instance of GIFT to be used and tested within a course's program of instruction, and the results of the pilot were promising with positive outcomes on assessments, as well as positive feedback from students about the GIFT system (Sinatra et al., 2022).

STEEL-R

The last use case reported in this chapter examines GIFT's role in an ecosystem paradigm that incorporates competency development, persistent data management, and builds from the Advanced Distributed Learning (ADL) Initiative's Total Learning Architecture (TLA; Walcott & Schatz, 2019). It extends the TLA by using experiential learning theory as its guiding construct (Kolb, 2004) and examines capabilities to drive longitudinal data capture with context for the purpose of long-term skill acquisition modeling. With these defined goals, the STE Experiential Learning for Readiness (STEEL-R; Goldberg et al., 2021) architecture and data strategy was created. STEEL-R integrated GIFT with key components and processes from the TLA (Walcutt & Schatz, 2019) and uses standards and data-science principles where feasible (Hernandez et al., 2022). This was primarily supported through the implementation of the xAPI (eXperience Application Programming Interface) data specification and developing an xAPI Profile linked to GIFT's DKF. This extension to GIFT enables automated production of formative and summative assessment statements in xAPI at the interaction, process and procedure levels, with careful attention for tracking when a skill is being applied and under what context that skill is being performed (Robson et al., 2022). The STEEL-R data strategy is being applied to build competency frameworks and profiles linked across cognitive, psychomotor, affective and team-oriented competencies, with a goal of explicitly defining what is required for an individual or team to be successful at their assigned job. The STEEL-R capability is under active research and development, and will serve as the foundational architecture to support future competency-based experiential learning research being managed at DEVCOM Soldier Center.

SWOT Analysis

As demonstrated in the section and timeline above, GIFT's domain-independent design has been influenced and guided over the years by several different topic areas and serves many different functions. However, the implementation methodology and program objectives also conversely provide several challenges. In order to assist in objectively evaluating GIFT's progress to date and framing how to improve its functionality, we have created the following self-assessed Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis for GIFT. In the following sub-sections, we define and summarize specific variables across the SWOT categories, followed by short narrative descriptions for each of the categories.

Overall SWOT Analysis

The following SWOT Analysis Table (Table 1) provides an overview for the reader, and are expanded upon in the section sub-headings below.

Table 1. Summary of Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis for GIFT

Strengths
GIFT is domain-independent, free and open-source
GIFT is flexible open-system architecture, designed for re-use, and adheres to evolving standards
GIFT is continuously evolving through research investments with regularly updated public releases
Provides data-enabled evidence of practice and adheres to xAPI specification
Incorporates Course Objects and Pedagogical Management based on ITS best practices
Multiple use cases across multiple learning domains to assist new users, with emphasis on serious games and dynamic simulations
Supports individual and team learning opportunities with customizable data outputs
Supports mobile technology content, interaction, and GPS tracking
GIFT Portal supports a large and diverse community of contributors
GIFT development has been guided by experts and practitioner communities: The Design Recommendations book series, Expert Workshops and GIFT User Symposiums
GIFT is well published and cited across high impact journals and conferences
GIFT comes in three versions, and the Cloud version does not require a download/installation for use.

Weaknesses
Limited Fielding and Data outside of experimental protocols
Limited automation linked to course authoring and DKF configuration; Re-design required to make it more user friendly
GIFT is a research tool; not all features and tools are fully developed and stable
Limited resources in program execution; Documentation not always current
Lack of Internationalization
Lack of self-contained user management system, requiring external user authentication
GIFT Course Ontology mismatches
Rigid domain logic linked to external training environments
Being a generalized framework limits specificity
Lack of mature data analytic tools and visualization dashboards
Challenging user-centered requirements

Opportunities
IEEE Standards Working Groups and AIS Consortium interest of GIFT in commercial sector
Evolution of 5G and future data networks and strategies
Outreach via GIFTSym and online conferences
Integration into commercial LMSs
Investment in Competency-Based Training across all DoD emphasizing need
GIFT interoperability with evolving Total Learning Architecture
Maturation and Adoption of eXtended Reality (XR) Learning and the Metaverse
Improvements in Artificial Intelligence (AI)
COVID-19 pandemic emphasized the need for adaptive computer-based learning technologies

Threats
Lack of adoption due to commercial competition
A growing user base with rapidly evolving needs
Uncertain long term funding strategy
Too many priorities, not enough time
Data and network security requirements
General privacy concerns
Instructor hesitation and the need for cultural buy-in

Strengths

GIFT is domain-independent: GIFT is designed to support multiple domains addressing cognitive, metacognitive, psychomotor, and affective learning objectives. The principle of domain-independence means that the domain module contains all of the domain-specific content.

Designed for individual and team learning opportunities: GIFT is also learner-independent in the sense that it provides mechanisms to model, assess, and influence both individuals and organizational team structures that drive the domain GIFT is being harnessed to support. These distinctions are explicitly highlighted within the use cases available to the public and under government purpose rights.

GIFT is flexible and serves as an open-system architecture: GIFT is flexible and can be applied to support multiple training environments and learning resources. GIFT is designed to operate under different configurations and across multiple data sources and types and supports individual and team tutoring requirements. In addition, GIFT can serve as an open system architecture to enable future extensions and support interoperability with evolving capabilities and data sources matured in industry or through other services and labs. This will promote longevity and enables an ability to leverage innovative capabilities that optimize current methods (e.g., integrating a new confusion classifier that will support real-time assessment, and providing new context to inform adaptive strategies) or retire current tools and methods that are obsolete.

Provides evidence of practice: GIFT collects logs, video, and audio files during a session that provide detailed information on the events recorded during a training interaction. GIFT also writes xAPI statements to a Learner Record Store (LRS). These artifacts populate a data set that can be used to determine competency and make other overarching generalizations of individuals and team, while providing data for tracking longitudinally (Hoffman & Goldberg, 2022).

Drives re-use: Enhancements to GIFT are shared with the community on a frequent basis through software releases and a hosted GIFT Cloud instance. GIFT utilizes standards (e.g., xAPI, DIS) and free/open-source platform data formats (e.g., JSON, Protobuf). GIFT also integrates with third party systems such as VBS3, Unity, Unity WebGL, TC3Sim, UrbanSim, Android/GPS, EST2, SAMT, and several more. Providing these plug-in solutions promotes re-use and extensibility.

Framework to implement ITS best practices: Within the goals of driving re-use, GIFT allows creation of custom course objects that support operationalized learning and best practice application. An example is the evolving Adaptive Course Flow object that builds on Merrill's Component Display Theory to sequence a flow of interaction, while incorporating Chi's ICAP framework (Chi & Wylie, 2014; Spain et al., 2019) to manage adaptive remediation selection. These approaches can be leveraged by practitioners and guide configurations to support these best practice approaches. As an aside, there are few, if any, applications that exist and are still used today where researchers and practitioners in ITSs can see what others are recommending, implementing in their own research, sharing with others, and leveraging a community backed initiative. GIFT provides a framework supported by different types of users, from Information System Developers (ISDs) to software engineers.

Multiple use cases across multiple learning domains: There are many different use cases that were developed (as demonstrated in the introduction above), which cover various topics. These use cases utilize a wide variety of environment types leveraging gaming, mixed reality, and mobile computing technologies. Each use case highlights one or more unique learning objective that is influenced by GIFT's ITS assessment and pedagogical functions.

Flexibility in data output: GIFT is researcher friendly. It provides a researcher the flexibility to output specific data types using the Event Report Tool. Raw log files are easily accessible, and data types and reporting structures are fully customizable using GIFT's source-code.

GIFT Supports and Adheres to Evolving Standards: As AIS standards continue to develop and become adopted, GIFT aligns with them as much as possible, and informs where feasible. This includes active participation on IEEE's Industry Consortium on Learning Engineering (ICICLE; IEEE, 2023), as well as participation on several working groups sponsored under their Learning Technology Standards Committee (LTSC), including Reusable Competency Modeling and Data Standards, Adaptive Instructional Systems, and Learning Object Metadata.

Online Access: The GIFT Portal (GIFTtutoring.org) enables a large and diverse user base, and includes an open forum for development and troubleshooting support. The portal also makes it easy to download the open source desktop version of the software. GIFTCloud (https://cloud.gifttutoring.org/dashboard/#login) is hosted by Amazon Web Services and provides access to the core tools, methods and workflows to build, deliver, and evaluate a GIFT authored lesson without requiring installs and system processes running on a local machine.

GIFT Development Guided by Experts and Practitioners: The Design Recommendations book series, Expert Workshops, and GIFT User Symposiums have been assisting in guiding GIFT development and improvements over the past decade. The GIFT User Symposium specifically tracks recommendations, projects, and contributions.

GIFT is well published in high impact journals and conferences: GIFT's development over time has been documented in the literature, and many of the studies conducted with GIFT have been published in journals and conferences.

Integration into the Army's Synthetic Training Environment (STE): GIFT will serve as the STE's ITS service, providing a multi-open system approach for managing multi-modal data capture, objectives assessments, feedback and coaching, adaptive injects and ITS-influenced after-action-reviews (AARs). This will theoretically establish a transition path to push maturing capabilities and research directly into the fielded solution. STE can also provide a rich data set through its fielding back to the research community to enable AI research methods reliant on big data.

GIFT is continuously updating and maintaining releases of software: There have been consistent yearly releases of a regression-tested desktop version of GIFT. The Cloud instance of GIFT is frequently updated as new functions and capabilities are developed.

GIFT supports Game-Based Training: GIFT provides tools and methods to create objective assessments in serious games and dynamic simulations. There are multiple examples that highlight the use of GIFT condition classes to convert game state information (e.g., DIS, HLA, Google Protobuf) into performance and behavior derived metrics that are used to assess performance. There are several examples across the use cases listed above, and documentation to assist developers in building new condition classes. As an open-system architecture, GIFT also supports interoperability with what we call External Assessment Engines.

GIFT is free and open-source: GIFT can be downloaded for free, or accessed via a free account online.

GIFT supports mobile learning and GPS tracking: GIFT works on mobile devices and GPS location has been demonstrated to be able to be utilized by the system. A mobile event course object is available to users, and provides the framework to create a GPS enabled learning experience configured to a physical location. A current use case is training Land Navigation procedures.

GIFT comes in three versions: GIFT has an installable desktop version, an easy to access cloud version, and an image you can run on your own server, enabling a controlled instance of GIFT Cloud managed under an organization's security and IT policy. GIFT Cloud works on multiple operating systems and does not require a download.

Weaknesses

Limited fielding and data: Most data collections are in support of developing and evaluating a specific functionality or pedagogical approach. There are limited 'big data' sets to date to properly train/validate models and policies. It is difficult to evaluate the effectiveness of GIFT in schoolhouses and training environments.

Limited automation linked to course authoring and DKF configuration: There are still considerable authoring and configuration requirements to enable a GIFT managed lesson with interactive and adaptive scenario content. This requires intimate knowledge of the GIFT features and workflows to take full advantage of its adaptive functions. To achieve use at scale, research is required to automate as much of the lesson creation and assessment configuration as possible.

Not all features and tools are fully developed and stable: As GIFT is primarily a research project, the features and research tools are developed based on need. Currently there are certain elements such as instructor dashboards and gradebooks that are not yet built. A features priorities list continues to evolve based on user needs in particular applications.

Limited resources in program execution; Documentation not always current: GIFT was built upon research vectors driven by Army directives. Compared to larger government acquisition programs and commercial applications, the GIFT team is small and limited in resources. This limitation prevents being able to fully explore every feature and use case. But our small size also makes us nimble enough to investigate more approaches and integration targets. Whereas documentation does exist for the features and processes in GIFT, it is not always updated when changes are made to the GIFT system. Maintenance of software systems is of course expensive. Consequently, there occasionally is outdated information.

Lack of internationalization: While a small effort is underway to support a Spanish translation of some of GIFT user interfaces (UIs), there needs to be a more dedicated approach in order to acquire users in other languages and promote GIFT as a true standard/best-practice. This involves better involvement across international professional societies, and building upon the relationships and collaborations across the Advanced Distributed Learning (ADL) Initiatives international landscape.

Lack of self-contained user management system: GIFT currently relies on GIFTtutoring.org as the user management system. This requires users to register on GIFTtutoring.org in order to authenticate when logging into the GIFT Dashboard. For GIFT installations that are located in secure or closed networks and those that are not affiliated with the core GIFT community, another user management system is needed. There is no best practice established in this arena.

GIFT course ontology mismatches: The current GIFT course ontology does not necessarily match what is found in the community. GIFT courses may be more traditionally considered as lessons. This difference in presentation and vocabulary may result in some confusion when first approaching GIFT. There is an opportunity to align the terminology of GIFT with more commonly used interpretations, and to make sure that the meaning of the terms used are clear.

Rigid domain logic linked to external training environments: Logic within some of the adaptive features (such as the Domain Knowledge File, DKF)) are very rigid. There is not much flexibility in the approaches for creating DKFs. There are potentially easier or more open ways to reach similar results.

Being a generalized framework limits specificity: Due to keeping the system as general as possible it sometimes makes it more difficult to design interfaces or fully represent all information that might be relevant to a specific domain. Translating strategies (i.e, general system actions) into tactics (i.e., specific implementation of a strategy) is required for each external training environment interfacing with GIFT.

Lack of mature data analytic tools and visualization dashboards: Data logs can be visualized and explored via the Game Master, but data extraction and analysis is currently limited to export tools and .csv files. There are no current visualizations of performance outputs that can be easily viewed and interpreted at the training objective and skill acquisition level.

Challenging user-centered requirements: GIFT has mostly focused on the researcher interfaces and there is no distinction between user roles. Students and instructors see the same interfaces when they login. There has been a recent effort to limit permissions to certain courses, but more work is needed on creating different interfaces based on the user type.

Opportunities

IEEE and AIS Consortium Adoption: An open-source version of GIFT will be shared through IEEE's Adaptive Instructional Systems (AIS) consortium, providing a mechanism for industry and academia to build and commercialize from a standard baseline. This will potentially provide a rapid evolution of capabilities that meet the training and education needs across multiple sectors, not just the Department of Defense (DoD).

Evolution of 5G and future networks: This advance removes any data bandwidth/latency issues preventing scalable solutions in the cloud. This is big for working with data-intensive training environments incorporating mixed reality (Virtual Reality and Augmented Reality), advanced intelligent sensing facilities, wearables, and a hybrid architecture approach.

Outreach via GIFTSym and online conferences: As many conferences have pivoted to online versions in the past few years there is an opportunity to reach additional individuals who may not have participated in an in-person version of the conferences. There is an opportunity for more people to engage with and learn about GIFT from online versions of GIFTSym.

Integration into commercial Learning Management Systems: GIFT would benefit from exploring and integrating with other platforms such as Learning Management Systems (LMSs). As has been the case in the past, integration leads to increased functionality and access to a larger user community.

Investment in competency-based training across DoD services: There is growing interest and attention in the area of competency-based training across all DoD services, with an emphasis on data-driven methods to track and influence training and education paradigms. This aligns with the goals and capabilities provided by GIFT, creating an opportunity to highlight GIFT's utility as a core Government Off the Shelf (GOTS) technology to facilitate meaningful data capture at the learning interaction level. In addition, there are evolving best practices for long term learning profiles that support the creation of longitudinal learner models with data that tracks performance over time.

GIFT interoperability with the Total Learning Architecture (TLA): Complimentary to the competency-based training opportunity, the advancement and adoption of the ADL's TLA enables GIFT to reach a broader audience by working as an assessment and tutoring service within the larger infrastructure. GIFT's xAPI profile establishes a data pipeline directly between a training interaction managed by GIFT and an enterprise instantiation of the TLA.

Maturation and adoption of eXtended Reality (XR) learning and the Metaverse: There is a wide adoption of virtual/augmented reality occurring in everyday life, in addition to opportunities regarding the metaverse. This provides great potential to provide engaging experiential learning opportunities in classroom and home settings. This proliferation creates an opportunity to embed intelligent tutoring functions to optimize these interactions, with GIFT's framework serving as an important starting point.

Improvements in Artificial Intelligence: GIFT will benefit from the serious maturation of Artificial Intelligence (AI) and is well positioned to benefit from AI open libraries. As an open system architecture, GIFT is designed to make use of evolving and maturing capabilities through its modular service-oriented design. As AI continues to advance in the areas of learning science, GIFT will be able to integrate and leverage those functionalities without significant engineering requirements.

COVID-19 and distributed learning emphasis: The COVID-19 pandemic emphasized a need for adaptive training and learning technologies. The pandemic also accelerated the timeline for the use of these technologies. This will accelerate the development and adoption of tools to better support learner needs in a distributed, self-regulated capacity.

Threats

Lack of adoption through commercial competition: There is a threat to large scale adoption based on commercial competition and building standards to promote industry benefit. As seen from the urgent need for at-home education and training made apparent during the COVID-19 pandemic, many companies increased their product capabilities in response. Another concern is a potential unwillingness to adopt a technology that was not developed in-house. There may be academic labs or businesses who prefer to create their technology themselves rather than using an existing software solution. Barriers to establish a standard centric framework for AIS evolution can be impacted as a result.

A growing user base with rapidly evolving needs: Designing for all users, current and future, to maintain concurrency and relevancy is a challenge. While GIFT and the GIFT authoring tools have intentionally been kept flexible to be used for many different uses, it can also lead to some confusion for authors. It is important to provide structure, supporting documents, and usability design that helps the author understand what needs to be completed for their specific use. If they get overwhelmed or confused, they may not use the system.

Uncertain long term funding strategy: While GIFT currently aligns with priorities and goals, there is not a guarantee that will always be the case. Sustainment models are not fully defined, and adoption outside of the government will be required to maintain its utility as an open framework to drive AIS utilization.

Too many priorities, not enough time: Aligned with evolving user needs, there are so many recommendations that are provided by the GIFT community, it is nearly impossible to reconcile all of them. Maturing a model to support community needs outside of the research investment space will be critical to scale and sustain the use of GIFT outside of the research investment.

Data and network security requirements: There are stringent security requirements for use in a military operational unit, which may make it difficult for GIFT to be adopted for use in military training. Investments are required to certify GIFT's core framework to work on secure networks, and to establish workflows to streamline this process accordingly.

General privacy concerns: Protecting user data is a critical necessity for any future enterprise level adaptive learning solution. Keeping current with data protection policy and requirements is critical to establish a GIFT capability that can evolve and support user data needs.

Instructor hesitation and cultural buy-in: An instructor may have a concern that robots and AI will eventually replace their job function. Creating a culture of buy-in is critical to the success of AIS use at scale. There are instructors and organizations that may be hesitant to adopt technology that could potentially replace traditional approaches.

Discussion and Conclusions

Our SWOT analysis of GIFT revealed a large number of strengths, weaknesses, opportunities and threats. There were many positive strengths identified within GIFT, primarily including its flexible nature (generalizability, open source, domain-independent), the documentation of its development in the literature and at the GIFT Portal, and its consistency (adopting Standards, having a release cycle, etc). While a number of weaknesses were identified, they tended to be items that as of yet have not been a primary focus of development, but have been noted by the GIFT team through the years, such as improving user interfaces, implementing user roles, improving authoring tools, maturing data visualization tools, and demonstrating GIFT in more domains. There were many opportunities that were identified which generally fell into the categories of: implementing emerging technologies, integrating with other systems, and incorporating improvements into the GIFT system (e.g., long term learner profiles, authoring tools). A number of threats were identified which include threats to adoption (e.g., commercial replacements, stringent security requirements, instructor hesitation), as well as threats to development (e.g., too many priorities, designing for all users, funding limitations). In some cases, work is currently addressing some of the gaps that exist within GIFT. In other cases, new gaps and opportunities for the future have been identified.

While many of the chapters contained within this book address elements of intelligent tutoring systems in general, the current chapter focused on the Generalized Intelligent Framework for Tutoring software. The process of completing this SWOT analysis helped to highlight the elements of GIFT that are most successful, as well as identified opportunities and areas for improvement. This analysis and the recommendations for GIFT that are associated with the additional chapters in this book will help provide a path forward as GIFT continues to develop.

Acknowledgements

The research described herein has been sponsored by the U.S Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. We would like to acknowledge Michael Hoffman for providing comments and initial edits on early versions of the chapter including contributing to the SWOT Analysis. We would like to acknowledge Elyse Burmester for contributing to the brainstorming for the SWOT Analysis.

References

Aleven, V., Sewall, J., Andres, JM., Sottilare, R., Long, Rodney & Baker, R. (2018). Towards adapting to learners at scale: integrating mooc and intelligent tutoring frameworks. In *Proceedings of the fifth annual ACM conference on learning at scale* (p.14). ACM.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52*(1), 1-26.

Biswas, G., Rajendran, R., Mohammed, N., Goldberg, B. S., Sottilare, R. A., Brawner, K., & Hoffman, M. (2019). Multilevel learner modeling in training environments for complex decision making. *IEEE Transactions on Learning Technologies*, *13*(1), 172-185.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307.

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

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., ... & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152-193.

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

Goldberg, B., Roberts, N., Powell, W.G. & Burmester, E. (2018). Intelligent Tutoring in the Wild: Leveraging Mobile App Technology to Guide Live Training. In *Proceedings of the International Defense and Homeland Security Simulation Workshop of the I3M Conference 2018*. Budapest, Hungary.

Goldberg, B., Amburn, C., Ragusa, C., & Chen, D. W. (2018). Modeling expert behavior in support of an adaptive psychomotor training environment: A marksmanship use case. *International Journal of Artificial Intelligence in Education*, *28*(2), 194-224.

Goldberg, B., Hoffman, M., & Tarr, R. (2015). Authoring instructional management logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). In R. Sottilare, A. Graesser, X. Hu, & K. Brawner (Eds.) *Design Recommendations for Intelligent Tutoring Systems, Vol. 3: Authoring Tools*, U.S. Army Research Laboratory.

Goldberg, B., & Cannon-Bowers, J. (2015). Feedback source modality effects on training outcomes in a serious game: Pedagogical agents make a difference. *Computers in Human Behavior*, *52*, 1-11.

Goldberg, B., Brawner, K., Sottilare, R., Tarr, R., Billings, D., & Malone, N. (2012). Use of Evidence-Based Strategies to Enhance the Extensibility of Adaptive Tutoring Technologies. In *Proceedings of the Interservice/Industry Training Systems & Education Conference*, Orlando, Florida, December 3 – 6, 2012.

Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of Educational Psychology, 89*(1), 92.

Hoffman, M. & Goldberg, B. (2022). The GIFT Architecture and Features Update: 2022 Edition. In *Proceedings of the 10th Annual GIFT Users Symposium* (pp. 11 – 20).

IEEE (2023). ICICLE – Learning Engineering. Retrieved on 23 January 2023 from https://sagroups.ieee.org/icicle/

IMS Global. (n.d.) Learning Tools Interoperability^{TM} Implementation Guilde (Final Version 1.3). Retrieved on 01 May 2022 from https://www.imsglobal.org/specs/ltiv1p1/implementation-guide.

Maidstone, R. (2012). Discrete event simulation, system dynamics and agent based simulation: Discussion and comparison. *System, 1*(6), 1-6.

Mall, H., & Goldberg, B. (2014). SIMILE: an authoring and reasoning system for GIFT. In *Proceedings of the 2nd Annual GIFT Users Symposium*.

McAlinden, R., Gordon, A. S., Lane, H. C., & Pynadath, D. (2008). UrbanSim: A game-based simulation for counterinsurgency and stability-focused operations. University of Southern California Los Angeles.

Ouverson, K. M., Ostrander, A. G., Walton, J., Kohl, A., Gilbert, S. B., Dorneich, M. C., ... & Sinatra, A. M. (2021). Analysis of Communication, Team Situational Awareness, and Feedback in a Three-Person Intelligent Team Tutoring System. *Frontiers in Psychology*, *12*.

Ostrander, A., Bonner, D., Walton, J., Slavina, A., Ouverson, K., Kohl, A., ... & Winer, E. (2020). Evaluation of an intelligent team tutoring system for a collaborative two-person problem: Surveillance. *Computers in Human Behavior*, *104*, 105873.

Robson, R., Ray, F., Hernandez, M., Blake-Plock, S., Casey, C., Hoyt, W., Owens, K., Hoffman, M. & Goldberg, B. (2022). Mining Artificially Generated Data to Estimate Competency. In *Proceedings of 2022 International Conference on Educational Data Mining (EDM)*. Durham, United Kingdom.

Sinatra, A.M. (Ed.). (2022). Proceedings of the 10th Annual GIFT Users Symposium. Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center. ISBN 978-0-9977258-2-7. Available at: https://www.gifttutoring.org/documents/159

Sinatra, A.M., Robinson, R.L., Goldberg, B., & Goodwin, G. (2022). Generalized Intelligent Framework for Tutoring (GIFT) Master Gunner Course Pilot Study. Virtual Poster presented at the 2022 Army University Learning Symposium, July 11th, 2022.

Sinatra, A. M., Sims, V. K., & Sottilare, R. A. (2014). The Impact of Need for Cognition and Self-Reference on Tutoring a Deductive Reasoning Skill. Army Research Lab, Aberdeen Proving Ground, MD.

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

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

Spain, R., Rowe, J., Goldberg, B., Pokorny, R., Lester, J., & Rockville, M. D. (2019, December). Enhancing learning outcomes through adaptive remediation with GIFT. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*. Paper (No. 19275).

Spain, R., Rowe, J., Smith, A., Goldberg, B., Pokorny, B., Mott, B. & Lester, J. (2021). A reinforcement learning approach to adaptive remediation in online learning. *Journal of Defense Modeling and Simulation (Special Issue on Artificial Intelligence in Education)*, *19* (2), 173-193.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: a meta-analysis. *Psychological Bulletin, 121*(3), 371.

Chapter 2 – Intelligent Tutoring Systems SWOT Analysis

Robert Sottilare[1] and Kurt VanLehn[2]
Soar Technology, Inc.[1]; Arizona State University[2]

Introduction

This chapter examines general trends for intelligent tutoring system (ITS) capabilities using the strengths, weaknesses, opportunities, and threats (SWOT) analysis methodology. ITSs are a type of computer-based training and education technology categorized as an adaptive instructional system (AIS) that accommodates individual differences (tailoring) to facilitate learner knowledge acquisition (Wang & Walberg, 1983; Tsai & Hsu, 2012) and to guide one-to-one learning activities that exercise skills defined by learning objectives (Sottilare & Brawner, 2018). ITSs use artificial intelligence (AI) and other advanced technologies to help people learn more effectively and efficiently (AIS Consortium, 2021).

Analysis Scope

The scope of our SWOT analysis is broad and considers emerging ITS technology, tools, and methods from both academic institutions and high-tech providers along with the state-of-practice commercial products. Given our SWOT analysis is about general trends across all ITS technologies (tools and methods), we use examples to highlight the state-of-practice and the state-of-the-art, but do not heavily focus on any single ITS technology or architecture (e.g., AutoTutor or the Cognitive Tutor). Our analysis also considers instances of how ITS technology is being used now, who uses it, how regularly they use it, where they use it, and for what specific purposes. Consideration is also given to the need for standards, and interoperability, accessibility, scalability, extensibility, maintainability, granularity of data, and usability trends in current and emerging ITS technology. In addition to learners (students), we also consider other users: course authors (creators), content curators, real-time learner monitors, and subject matter experts (SMEs) also known as domain knowledge providers. Finally, in our analysis we embrace the idea that weaknesses and threats can be distilled down into opportunities.

Strengths

While the attributes in this section are not present in all ITSs, the strengths described below are present in a majority of ITSs and are considered to have advantages over previous technologies, tools, or methods:

Effectiveness: Some ITSs have been empirically demonstrated to be as effective as expert tutors (VanLehn, 2011). Kulik and Fletcher (2016) conducted a meta-analysis of findings from 50 controlled evaluations of ITSs and the median effect in the 50 evaluations reviewed raised test scores 0.66 standard deviations over conventional methods (e.g., traditional classroom training, an increase from the 50th to the 75th percentile).

Engagement: ITSs, when paired with virtual characters or game-based training environments to support one-to-one tutoring, can be much more engaging than classroom training.

Granularity: ITS are often implemented with user interfaces that collect fine-grained data on user performance. The advantage of granular data is that it can be manipulated (e.g., aggregated or disaggregated) to support adaptation in a variety of conditions and situations. In particular, ITSs can provide fine-grained feedback and hints.

Personalization: ITSs can tailor their actions to the learner's needs. ITSs can change what they say and do depending on the learner's performance, workload, emotional states or other individual differences.

Accelerated learning (Efficiency): The personalization features in ITSs also provide the opportunity to reduce the time required for a learner to reach proficiency because most of the instructional contact time with learners is focused on learning gaps instead of prior knowledge.

Cognitive domains: ITSs, commercial and academic, have primarily focused on the cognitive domain to support problem solving, decision making and procedural tasks bringing a higher level of understanding about the nature of cognitive domains compared to community experience with other domains (affective, psychomotor or collaborative).

Recommender systems: To guide learners in their quest to increase their knowledge and skills, ITSs are designed to continuously track their progress toward goals and recommend/select content needed to reduce learning gaps (differences between learner competency and established learning objectives).

Pairing ITSs with other technologies: ITSs are being paired with virtual, augmented, and mixed reality (XR) technologies to support training domains that require more visual stimuli (Gilbert, Intelligent Tutoring System PADLET, 2021). There are good examples of successful implementations that link ITSs with game-based training environments to support adaptive instruction.

Weaknesses

While the attributes in this section are not present in all ITSs, the weaknesses described below are present in a majority of ITSs and are negative drivers of ITS cost, efficiency, and performance:

Authoring systems: ITS authoring systems used to create adaptive instructional courses are often complex and require expert knowledge including instructional design, domain knowledge, or computer programming skills (Sottilare, Intelligent Tutoring System PADLET, 2021). The systems often have complex procedures and do not provide necessary guidance to complete a functional course.

Non-cognitive domains: ITS architectures have primarily focused on the tutoring of cognitive domains and have largely ignored affective (value or ethic-based) domains, psychomotor domains, and collaborative domains of instruction. GIFT and a few ITSs have recently begun the process of addressing architectural requirements for designing, authoring, deploying, and evaluating the effect of instruction in these domains. Slowly, more diverse and complex domains are being represented in adaptive courses (Sottilare, Intelligent Tutoring Systems PADLET, 2021).

Multi-modal ITSs: Many ITSs provide some prescriptive interaction between learners and the tutor. This interaction is highly constrained (e.g., multiple choice inputs and responses) and mostly text based (Sottilare, Intelligent Tutoring System PADLET, 2021). While AutoTutor remains the primary shining example of a conversational tutor, a growing number of ITSs are integrating virtual character frameworks (e.g., University of Southern California – Institute for Creative Technologies' Virtual Human Toolkit) that provide both an embodied conversational agent and the logic required for natural language understanding and generation of appropriate responses. The ability to integrate virtual human frameworks with ITSs will be essential to providing mixed initiative dialogue at scale.

Measures of ITS effectiveness: It is difficult to evaluate ITS effectiveness in terms of their ability to influence learner performance (Sottilare, Intelligent Tutoring System PADLET, 2021). ITS performance is often based on changes in learner performance and compared to traditional classroom training as a baseline. It is important to be able to produce an apple-to-apple comparison of system performance and some analyses have skewed effectiveness results by also including the impact of improved content. For example, a study (Fletcher, 2011) reported an effect size of 2.81 sigma in comparing a digital tutor and an integrated learning environment teaching information systems technology. However, part of the published impact was due to a refreshing of content and not entirely due to the adaptive capabilities of the tutor. We must also be consistent in describing learner behaviors and their relationship to learning outcomes. For example, mapping user clickstreams in the user interface is important to understanding learner behaviors during the adaptive instructional process. It is important to note that while there may be a relationship between learner behaviors and available datastreams, it might not be useful as a measure of learner performance or ITS effectiveness (Gilbert, Intelligent Tutoring System PADLET, 2021).

Development & maintenance costs and return-on-investment (ROI): ITSs are complex systems that require expert knowledge to design, develop, test, deploy, and evaluate. ITSs used by large populations may demonstrate a sufficient ROI to merit the investment, but specialized ITSs used by small learner populations often fail to provide a sufficient ROI (Fletcher & Sottilare, 2014). How can we reduce development time and costs to encourage community investment in ITSs? Automation may hold the key to reducing development and maintenance costs (total ownership costs).

Accessibility: ITSs are designed to be compatible with common internet browsers (e.g., Chrome, Safari, Edge) and are often accessible from laptops, workstations, tablets, and smartphones. While this might be perceived as a strength, ITSs designed to operate with heavy computation loads (e.g., machine learning algorithms to classify learner workload or other states) may not be usable on smartphones that generally have a low computational capacity. This computational limitation may limit accessibility to ITSs in marginal populations where laptops and workstations are not widely available. ITSs are often not designed to easily accommodate disabled students, which can thwart adoption by some institutions. ITSs are often not designed for use in languages other than the author's language.

Lack of fine-grained personalization: ITSs can personalize by selecting tasks but rarely by selecting appropriate messages on the step level. The technology is sufficient to support fine-grained personalization, but the theories of how to adapt feedback, help messages or other interventions to individuals are lacking.

Lack of mixed initiative dialogue (one-shot communication): If learners do not understand a feedback or help message, most ITSs do not permit learners to ask questions about it and get an on-point response from the ITS. Confused learners often ask confusing questions.

Lack of learner control: ITSs are generally designed to guide learning experiences and control the process of adaptive instruction. Areas where learners might be given more control include controlling the choice of a virtual instructor's appearance and interaction, controlling the choice of learning peers, initiation and selection of on-demand learning topics, control over their learner model and their information, control over the domain and teaching approach, and control over the amount of control the learner has over ITS processes (Kay, 2001).

Sustainability: ITS are often designed for powerful performance rather than easy maintenance or component reuse. ITSs are often poorly documented, making it difficult to pass security scans required by many institutions of their courseware. The ability to reuse components from one ITS in another ITS is currently nearly 0% due to the lack of interface and data standards.

Human instructor involvement: Human instructors can watch the tutoring unfold, but have little control over it at the step level. Unless they were involved in developing the ITS, they may disagree with the ITS's interactions with the learner, and may be limited by their ability to alter ITS interactions. The ability for teachers to have some influence over ITS interactions with their students may go a long way in building their trust of ITSs and other education technology.

Speech, groups and teams during collaborative learning: Learners interacting with each other in small groups or teams often leave the ITS out due to its limited understanding of their speech and the inability to generate appropriate responses.

Lack of bonding: Students often develop a bond with human teachers that they do not develop with ITSs, and this impacts their compliance and motivation. For example, a teacher can successfully ask students to collaborate, but students may ignore an ITS that asks them to collaborate.

Prescriptive (less flexible) systems: A defined domain should have well-written and measurable learning objectives. A well-defined objective for a geology course should describe a learning outcome (e.g., the student will be able to distinguish between igneous and sedimentary rock samples), should be learner-oriented, and be observable (or describe an observable product). Ill-defined domains (e.g., law and medical diagnosis) require some interpretation of facts/information by the learner to successfully achieve a learning objective. Achieving an objective in an ill-defined domain may be one of many successful paths or outcomes.

While there has been a heavy concentration on the development of ITSs in well-defined domains such as computer programming and mathematics (Sottilare, Intelligent Tutoring Systems PADLET, 2021), there has also been ITS applications in science domains, reading comprehension, language learning and other domains that are not entirely well-defined (Graesser, Intelligent Tutoring System PADLET, 2021). ITSs are currently not well suited to operate in ill-defined domains. They generally require rules or heuristics to assess learner performance and these rules are usually the result of decades or even centuries of study.

Opportunities

The opportunities listed in this section are challenges that have not yet been addressed at scale and are considered to have high positive impact once viable solutions are mainstream in the marketplace:

Non-cognitive domains: Opportunities could be created with new markets for adaptive training in non-cognitive domains. GIFT and a few ITS frameworks have recently begun the process of addressing architectural requirements for designing, authoring, deploying, and evaluating the effect of instruction in these domains. Slowly, more diverse and complex domains are being represented in adaptive courses (Sottilare, Intelligent Tutoring Systems PADLET, 2021). The creation of new architectures to support the design, authoring, deployment, and evaluation of affective, psychomotor, and collaborative courses could extend and exploit existing markets. Opportunities exist in institutional training (e.g., business ethics), sports (e.g., golf) and military team training (e.g., Army squads or Naval combat information center operations).

Improved learner engagement: There are opportunities to improve learner engagement during interactions with ITSs. The improved realism and responsiveness of virtual humans and their integration with ITSs may enable bonding between ITSs and learners that has been lacking. Improved control over the attributes of virtual humans representing instructors and learning peers is also likely to improve learner engagement.

Self-improving systems: The ability of ITSs to improve their performance with each experience will mean better predictive accuracy of learner states and better decisions in selecting/crafting learner interventions (Sottilare, Intelligent Tutoring System PADLET, 2021). Reinforcement learning might be helpful in tailoring feedback and help messages as well as personalizing recommendations for future learning experiences. It might also be useful to consider how ITS users perceive their adaptive instructional experiences and use this information to rank content, interactions, and overall experiences to improve ITSs through an evolutionary process where the best content, best methods, and best interactions rise to the top.

Improved speech understanding: The ability to understand the natural language of multi-sided conversations will make ITSs more valuable as they participate in and guide the learning of groups and teams.

Analysis of learner questions: Collecting data on learner's questions about feedback and help messages might lead to more conversational interaction and better communication.

Easy to use authoring tools: Easy authoring processes that empower subject matter experts to build ITSs would improve ITS affordability and help proliferate their use. Four recommendations for new authoring features include: 1) automation for guided authoring and content curation processes, 2) authoring on lighter, more affordable platforms (tablets and smartphones), 3) automated authoring of after-action reviews (AARs) to recap adaptive learning experiences, and 4) knowledge management tools for adaptive course developers (Sottilare, Intelligent Tutoring System PADLET, 2021).

Authoring tool standards: Often, ITS authoring tools have a high level of system or tutoring engine specificity (McCarthy, Authoring Tools PADLET, 2021). Authoring may be unique to the system, tool or framework, and the instructional and learning theories prescribed by that system. Standards or at least recommended practices for ITS authoring processes could be useful in streamlining authoring processes (Graesser, Authoring Tools PADLET, 2021). Standards might also be useful in easing the transfer of ITS courses and components from one framework to another.

Interoperability Standards: Common interface and data standards for ITSs will provide an opportunity to reuse courses, components, models (learner, team, instructional, domain, and interface) and subsystems from one ITS in another. In 2019, the Adaptive Instructional Systems (AIS) Working Group under the Learning Technology Standards Committee was formed to support IEEE Project 2247 standards and recommended practices for ITSs, recommender systems and other types of AISs. The goal of the AIS Working Group is to model the AIS, its components, and data exchange mechanisms, define interoperability standards and recommended practices for AIS buyers to evaluate systems and support the ethical use of AI in adaptive instruction. Among the markets with large training infrastructure investments, AIS interoperability standards could enable the augmentation of existing training systems by adaptive instructional logic. This would alleviate the need to replace or totally redesign existing systems to take advantage of the features of AISs.

Multi-modal ITS: While AutoTutor remains the primary shining example of a conversational tutor, a growing number of ITSs are integrating virtual character frameworks (e.g., University of Southern California – Institute for Creative Technologies' Virtual Human Toolkit) that provide both an embodied conversational agent and the logic required for natural language understanding and generation of appropriate responses. The ability to integrate virtual human frameworks with ITSs will be essential to providing mixed initiative dialogue at scale.

Supporting ill-defined domains: We recommend two approaches to support ITS development in domains that are not fully defined. The first approach is to integrate AI that enables the ITS to learn from its own previous performance so that its decisions improve with experience. Ideally, in a large learner population there will be large datasets representing outcomes and the conditions of the learner (or team) and the environment that can be used to train and optimize ITS recognition of events and learner states along with instructional decisions (e.g., interventions involving changes to content difficulty and interactions with the learner). The second approach is to research and develop methods that can test plausible root causes of performance outcomes in sparse data environments. Most ITSs focus on identifying errors and this usually requires well-defined knowledge of the domain. Focusing on root causes will allow ITSs to adapt strategies to get better outcomes and eliminate barriers to learning instead of dealing with symptoms of poor performance. Approaches to root cause analysis include hypothesis testing methods (e.g., abductive reasoning) that identify the factors (e.g., learner or environmental conditions) contributing to learner performance.

Global investments: Opportunities for innovative education technology such as ITSs will continue to grow. According to Grand View Research (2021), the global markets for online education, artificial intelligence in education, and smart education & learning (all relevant to the ITS marketplace) are forecasted to grow at a compounded annual growth rate (CAGR) of 9.23%, 32.9%, and 17.9% respectively over the next 5-10 years.

Threats

The threats listed in this section are challenges that have not yet been addressed at scale and are considered to have high negative impact if no viable alternatives or solutions are developed in the next 3-5 years:

Adoption of ITS technology: As user communities search for viable education technology solutions, the adoption of ITS technology is limited by ITS effectiveness, efficiency, engagement, culture, and affordability (Sottilare, Intelligent Tutoring System PADLET, 2021). The effectiveness and efficiency of ITSs in delivering adaptive instruction is well documented, but evaluation methods for the marketplace to consider, compare, and contrast ITS capabilities are lacking. An inability to fully understand the salient characteristics and features of ITSs in the marketplace threatens their widespread use in domains outside of mathematics, physics and computer programming where there is less of a track record of use. Engagement has always been an important aspect of learning and will continue to drive some buying decisions in the ITS marketplace. For example, a tutor with more visual appeal may be selected over a system with more learning impact. Culture can take many forms, but trust is an important component of culture.

Culture does not necessarily influence how much we trust, but does influence the way we trust. Since ITSs have a significant AI element within their design and AI is often viewed with skepticism and misunderstanding, ITSs can also be viewed as mysterious and even threatening. Moreover, instructors' professional identity can be threatened by adding a second "teacher" to their classroom. ITS providers should consider their messaging and be transparent in how their products work to build trust in various communities. They should be active in supporting the ethical use of AI in their products and develop objective measures to help buyers distinguish between the ITS features in their products and other vendor products.

Affordability plays a role in the effectiveness of capabilities available to different cultures. Affordability also directly limits access. According to Alsop (2021), only 7.7 percent of households in Africa were estimated to have access to a computer at home, but O'Dea (2020) reports the percentage of smartphone owners is about 40% across sub-Saharan African countries. It will be important for ITS providers to understand how their users will be able to access ITS capabilities and adapt their capabilities to optimize user access.

We cannot end our discussion of culture without touching on system internationalization (Gilbert, Intelligent Tutoring System PADLET, 2021). System internationalization is the design and development of a product, application, or document content so that it can be localized/transformed for target audiences that vary in culture, region, or language. It goes beyond replacing English with, for example, French. This type of data-driven approach will help lower the barriers to ITS acceptance.

Overuse of hints and other ITS support strategies: Some ITSs use hints too frequently which can detract from instead of enable learning (Graesser, Intelligent Tutoring System PADLET, 2021; Sottilare, Intelligent Tutoring System PADLET, 2021). According to Durlach and Spain (2014), hints along with cues and prompts are methods used to provide support to learners during instruction (adaptive or otherwise). If enabled, some student abuse hints, game the system and invent strategies for evading learning and avoiding failure (Bell, Nye & Kelsey, 2019). To support a learner efficiently, "a teacher should predict how much support a learner must have to complete tasks and then decide the optimal degree of assistance to support the learner's development" (Ueno & Miyazawa, 2017, p. 415). Abuse of support strategies leads to hollow learning experiences where the learner may be assessed as proficient, but fails to achieve any deep learning in the domain of instruction.

Cohort cohesion: The practice of keeping groups of learners together for a defined period is often at odds with the concept of personalization which allows fast learners to master content before slow learners. The motivation to master content early is low in a cohort unless there is additional content to support learning at various levels of achievement (e.g., expert learner, mastery learner, proficient learner, basic learner). While cohort cohesion is not specific to adaptive instruction, it is exacerbated by personalized learning.

Low technology acceptance: Teachers have concerns that ITSs and other education technologies will cost them their job (Sottilare, Intelligent Tutoring System PADLET, 2021). There is anxiety about the efficiency and effectiveness of ITSs and whether improved ITSs could mean that human instructors might not have a future in classrooms. Decision makers for ITS purchases have concerns that ITS effectiveness does not justify the cost (Sottilare, Intelligent Tutoring System PADLET, 2021). Buyers and users are concerned that there are not enough low-cost options in the education technology marketplace (Sottilare, Intelligent Tutoring System PADLET, 2021).

The Adaptive Instructional Systems (AIS) Consortium (www.aisconsortium.org) advocates for education technologies such as ITSs and intelligent mentors. They see AISs as tools for use by teachers that will help them manage heavy workloads in large classes and make them more productive. The AIS Consortium stands for the ethical use of ITSs (and AI in education and training) to help students learn. The AIS Consortium does not see ITSs as a credible threat to replace human instructors.

Academic freedom: The freedom to continuously improve their teaching to fit their personality and experience are perceived by some United States (US) teachers as a core benefit of their job and professional identity. They view ITSs, or any other educational software that does not allow them to change lessons to their satisfaction, as a limitation to their authority in the classroom. Standardization of lessons might be acceptable in China and Europe, but are not widely accepted in US public schools.

Standardized certifications versus personalization: Although making a high-stakes hiring decision warrants delving into the candidates' personal attributes, making a low-stakes hiring decision (e.g., hiring a plumber to fix a leak) requires standard certificates. The potential for individually-tailored outcomes enabled by the personalization of ITSs may be perceived as a small demand in some markets, but global projections show steady double digit growth opportunities in the next 5-10 years (Grand View Research; https://www.grandviewresearch.com/press-release/global-education-technology-market).

SWOT Analysis Recap

The tables in this section provide an overview of the ITS SWOT analysis discussed in the previous section of this chapter. We have organized recommendations for future actions into six areas to highlight opportunities and reduce both ITS weaknesses and threats:

- Improving authoring and curation tools and methods (Table 1)
- Improving real-time and long-term performance (Table 2)
- Improving the accuracy of learner and team models (Table 3)
- Improving domain modeling and assessment processes (Table 4)
- Improving instructional strategies (Table 5)
- Improving ITS interfaces (Table 6)
- Improving the fit of ITSs into existing educational cultures (Table 7)

It is important to note that not all the recommendations provided are technology focused. Some are policy or standards developments that build confidence and trust for ITS technologies and solutions. It is also important to understand that the recommendations being made in the opportunities column of each table may not represent the need for totally new features. While these capabilities exist in some ITSs, they are not widespread nor are they standard features within every ITS as we are recommending.

Table 1. Recommendations for improving authoring and curation tools and methods

Weaknesses & Threats	Opportunities
Low usability of ITS authoring systems and processes	• Improve automation and guidance for authoring processes • Improve automation of content curation processes • Improve author knowledge management tools • Enable authoring of effective adaptive courses without knowledge of computer programming or instructional design • Enable easy integration of capabilities that increase the engagement, efficiency, and effectiveness of ITSs (e.g., integration of virtual humans and existing instructional infrastructure) • Develop ITS authoring standards for interfaces and processes to enhance transfer of authoring knowledge from one ITS to another
Limited accessibility	• Improve ITS accessibility by extending authoring processes to support adaptive course creation for tablets & smartphones

	● Extend the types of ITS-compatible browsers to improve accessibility ● Facilitate internationalization ● Facilitate use by those with physical impairments.
Limited concentration of ITSs in non-cognitive domains of instruction	● Authoring processes and templates for non-cognitive domains of instruction (affective, psychomotor, and collaborative)
High ITS authoring costs	● Reduce ITS authoring costs by reducing the skills required to author ITSs ● Reduce ITS authoring costs by promoting reuse through ITS interoperability standards

Table 2. Recommendations for improving real-time and long-term domain performance

Weaknesses & Threats	**Opportunities**
Lack of bonding with students	● Make it easy to integrate realistic and responsive virtual humans into ITSs in the roles of both instructors and peers ● Enable learners to control the appearance and other characteristics of virtual humans in ITSs
Communications limitations	● Improve natural language understanding and generation in ITSs ● Improve multi-modal communications in ITSs ● Improve the ability for human instructors to influence ITS communications ● Improve the communications capabilities of ITSs during collaborative interactions
Learner control limitations	● Enable mixed initiative dialogue to improve learner control (e.g., learner question asking) ● Enable more control choices for learners (e.g., choice of instructor and peers, ability to initiate dialogue and selection of on-demand learning topics)
Overuse of support strategies	● Use experimentation and reinforcement learning to determine better policies for offering and withholding support.

	● Develop better methods for detecting student gaming and other abuse of support strategies.

Table 3. Recommendations for improving the accuracy of learner and team models

Weaknesses & Threats	**Opportunities**
Inaccurate models of individual learners and teams limit the effectiveness of ITSs	● Enable machine learning classification and predictive analysis of learner states in both dense and sparse data environments ● Improve ITS accuracy and performance by designing ITSs as self-improving systems ● Design ITSs to allow for flexible learner and team modeling based on domain, measures of assessment and data sources (e.g., sensors) ● Design ITSs to enable identification of root causes of learner performance (Sottilare & Hoehn, 2021)

Table 4. Recommendations for improving domain modeling and assessment processes

Weaknesses & Threats	**Opportunities**
Lack of interoperability standards for domain models and assessment processes	● Develop interoperability standards and recommended practices that enable portability of domain models, expert models, and assessment standards across ITS platforms ● Develop interoperability standards for verbal and non-verbal communications
Lack of visualization for domain progress with respect to learning objectives	● Provide visualization in the form of an open learner model that highlights learner progress toward learning objectives ● Provide visualization of individual learners with respect to various segments of the learner population (e.g., classroom, school, all 9th grade algebra students)
Lack ability to support more ill-defined learning domains	● Provide the ability for ITS developers to author effective tutors in ill-defined domains (e.g., law, medical assessments) without defining every possible outcome ● Provide automated processes to feed ITS knowledge acquisition and assessments

Table 5. Recommendations for improving instructional strategies

Weaknesses & Threats	Opportunities
Lack of evidence-based methods to model the effectiveness of various instructional strategies (e.g., mastery learning, metacognitive strategies, content selection strategies)	● Develop evidence-based methods to model the effectiveness of various instructional strategies under varying learner and course conditions ● Develop and maintain long-term learner models to track the impact of instructional strategies within individuals and populations

Table 6. Recommendations for improving ITS interfaces

Weaknesses & Threats	Opportunities
Lack of standards for ITS interfaces	● Develop interface standards for integrating discrete event training simulations ● Develop interface standards for integrating virtual humans with ITSs ● Develop interface standards for integrating physiological and behavior sensors with ITSs ● Develop interface standards for learner-tutor verbal communications

Table 7. Recommendations for improving the fit of ITSs to existing educational cultures

Weaknesses & Threats	Opportunities
Adoption of ITS technology	● Large organizations (e.g., military training researchers & developers, education-related societies such as AI in Education, and the AIS Consortium) with significant investments in ITS tools and methods should continuously address concerns that limit their adoption.
Low technology acceptance	● Develop an ability of an ITS to explain its pedagogical decisions and policies to users, both in general and as they apply to specific episodes of learner interaction with the ITS. ● Understand and develop evidence-based modeling of ITS users and their populations, which should help reduce acceptance barriers and increase ITS adoption in current low use communities (e.g., low income or low-tech

	marginalized groups)
Academic freedom to customize instruction	● ITS flexibility might be greatly improved, but this could also be at the expense of usability. To balance flexibility and usability we recommend high usability for novice authors/instructors and maximum flexibility for more knowledgeable ITS authors/instructors so these users can tailor ITS features to their specific needs.
Standardization certification versus personalization	● While learners may be able to learn a personalized set of competencies, we recommend that they be advised about what competencies are required for certificates.

Discussion and Recommendations for Future Research

In addition to the recommendations provided in Tables 1-7, there will continue to be a need for modeling the interaction between learners and ITSs to understand the effect of ITS actions on individual learners and teams. While the processes for modeling learners are mature, there is still room to improve the modeling of both individual learners and teams in context, and then understand the most effective way to use this information to optimize ITS actions. System flexibility will also continue to be an important aspect of ITS design. Individual user (learners, instructors, authors) differences along with local policies will require robust ITS architectures that can enable both novice and expert users. The ability to automate or guide processes will make ITSs more beneficial and affordable. Finally, standards and recommended practices will guide ITS designers and developers, and enable higher degrees of reuse.

Recommendations for GIFT Overall

Based on our findings, we have developed a set of general recommendations for ITSs (Tables 1-7). In this section, we provide specific recommendations for the Generalized Intelligent Framework for Tutoring (GIFT).

Usability: How can we make it easier for subject matter experts to author adaptive courses in GIFT? First, we can exploit AI methods to guide authors in the development of GIFT courses. The development of a course creator status view will help authors with authoring tasks including content curation (search, retrieve, tagging, and storage of content), traceability of learner activities to learning objectives, assignment and tracking of measures of assessment, and development of domain-dependent learner interventions.

Standards: GIFT is noted for various architectural principles that have made GIFT a de facto standard for AIS interoperability and as a largely domain independent framework. The researchers and developers of GIFT have been diligent in sharing their thoughts about its design, development and experimentation in the literature and the use of GIFT in various domains of instruction. GIFT has been integrated with various training capabilities including simulation and game-based training systems (also known as serious games). This record of using GIFT as an exemplar to overcome various AIS limitations over the last 10 years has cemented its place as a leading AIS framework. IEEE AIS standards activity under Project 2247 continues to use GIFT as a basis for discussion in developing AIS models, interoperability standards, recommended practices for evaluation and the ethical use of AI in education technology. To continue this legacy, we recommend that GIFT be used as an experimental testbed to evaluate desired outcomes in the AIS marketplace.

Transfer of GIFT Principles: There is a perception that GIFT is strongly associated with the US Army and Department of Defense (Robson, Intelligent Tutoring System PADLET, 2021). Are there plans to promote GIFT to be used more broadly? The AIS Consortium has recently negotiated with the US Army to transition the GIFT 2021-2 version for open-source and commercial use as the Global Learning Toolkit (GLT; aisconsortium.cloud). The GLT as a forked baseline of GIFT went online in October 2021. The AIS Consortium now provides the GLT as an open-source ITS architecture to any user worldwide and commercial entities within the AIS Consortium plan to extend GLT with new services offered either as open-source or commercial plug-ins. The ability to use GLT will enable learning and instructional principles in GIFT to be analyzed by a larger user population resulting in modifications and more robust ITS capabilities.

Conclusions

ITS performance (decision-making) and accuracy may be greatly improved through automation and built-in guidance that tracks the progress of both adaptive course authors and learners. Methods are needed to be able to support ITSs processes in both dense and sparse data environments. Standards will enable more flexibility for buyers and users, promote reuse, and help build trust in ITS technologies. ITS development costs may be reduced over time as reuse increases and the skill required to create and use ITSs decreases.

References

AIS Consortium (2021). Adaptive Instructional Systems defined. In the Charter of the Adaptive Instructional Systems (AIS) Consortium. Approved 6 January 2021.

Alsop, T. (2021). Computer penetration among households in Africa 2005-2019. Statista. Retrieved on 23 April 2022 from: https://www.statista.com/aboutus/our-research-commitment/2474/thomas--alsop

Bell, B., Nye, B., & Kelsey, E. (2019). Toward a Generalized Appliance for Measuring Engagement and Motivation Across Learning Environments. ITEC 2019.

Durlach, P. J., & Spain, R. D. (2014). Framework for instructional technology: Methods of implementing adaptive training and education. *ARMY RESEARCH INST FOR THE BEHAVIORAL AND SOCIAL SCIENCES*, FORT BELVOIR VA.

Fletcher, J.D. (2011). DARPA Education Dominance Program: April 2010 and November 2010 *Digital Tutor Assessments*.

Fletcher, J., & Sottilare, R. (2014). Cost Analysis for Training & Educational Systems. Design recommendations for intelligent tutoring systems, 2.

Gilbert, S. (2021). *Intelligent Tutoring Systems PADLET Comment*

Graesser, A. (2021). *Authoring Tools PADLET Comment*

Grand View Research. (2021). Education Technology Market Size Worth $377.85 Billion By 2028. https://www.grandviewresearch.com/press-release/global-education-technology-market

Kay, J. (2001). Learner control. *User modeling and user-adapted interaction*, 11(1), 111-127.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, 86(1), 42-78.

McCarthy, J. (2021). *Authoring Tools PADLET Comment*

O'Dea, S. (2020). Smartphone users in South Africa 2014-2023. Statista. Retrieved on 23 April 2022 from: https://www.statista.com/statistics/488376/forecast-of-smartphone-users-in-south-africa/

Robson, R. (2021). *Intelligent Tutoring Systems PADLET Comment*.

Sottilare, R. (2021). *Intelligent Tutoring Systems PADLET Comment*

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 *14th*

International Conference of the Intelligent Tutoring Systems (ITS), Montreal, Quebec, Canada.

Sottilare, R. & Hoehn, R. (2021). Investigation of AI Methods to Support Accurate Root Cause Analysis of Learner States - Research summary report for the Learner Data Institute.

Tsai CC., Hsu CY. (2012) Adaptive Instruction Systems and Learning. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_1092

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

Vygotsky, L. (1987). Zone of proximal development. Mind in society: The development of higher psychological processes, 5291, 157.

Wang, M. C., & Walberg, H. J. (1983). Adaptive instruction and classroom time. *American Educational Research Journal*, 20(4), 601-626.

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

SECTION II– INTELLIGENT TUTORING SYSTEM COMPONENTS SWOT ANALYSES

SWOT Analyses of:

Learner Modeling

Instructional Strategies

Authoring Tools

Domain Modeling

Chapter 3 - Learner Modeling in Intelligent Tutoring Systems SWOT Analysis

James Lester, Anisha Gupta, Fahmid Morshed Fahid, and Jay Pande
North Carolina State University

Introduction

Learner modeling has long been a central functionality of adaptive learning environments. Because robust learner models can drive adaptivity, the field of AI in education has been engaged in a decades-long exploration of learner modeling. Research on learner models has traditionally been concerned with representing and inferring learner knowledge components and skills competencies (Pelánek, 2017; VanLehn, 1988; Yudelson et al., 2013). Classic research on learner modeling has ranged from models of learners' knowledge and skills (Pelánek, 2017) to models of learners' plans, goals, and preferences (Chrysafiadi & Virvou, 2013), and recent years have seen the emergence of increasingly powerful inference methods (Gupta et al., 2021; Putra et al., 2021).

In this chapter, we present a SWOT analysis of learner modeling. First, we discuss the strengths of learner modeling produced by 50 years of advances in the field. Next, we turn to weaknesses, which have arisen in large part because of the field's historically relatively narrow focus. We then move to a discussion of opportunities presented by advances in underlying technologies. We next discuss threats that will be important to address, particularly considering the increasing adoption of AI-driven learning technologies in a wide range of education and training settings. Finally, we turn to the future of learner modeling, where increasingly accurate learner models will inform a broad range of pedagogical adaptations.

SWOT Analysis

Strengths

Learner modeling research is extraordinarily robust. State-of-the-art learner modeling functionalities are remarkably robust (Biswas et al., 2019; Chrysafiadi and Virvou, 2013; Owen et al., 2019; Shute et al., 2021). There is a rich history of learner modeling that began in the 1970s (Wenger, 1987). The earliest intelligent tutoring systems had primitive overlay models (Carr & Goldstein, 1977), and over the course of the evolution of AI learning technologies, probabilistic learner modeling techniques, such as Bayesian Knowledge Tracing (Gervet, 2020; Pelánek, 2017; Yudelson et al., 2013), have become increasingly prevalent. Learner modeling has become a cottage industry in the Artificial Intelligence in Education (AIED), educational data mining, and learning analytics communities, and there is now an enormous literature on learner modeling.

Historically, learner modeling has been notably strong for well-defined domains. Because of the relatively straightforward representational requirements of subject matters such as mathematics, physics, and, to some extent, computer science and computational thinking, there was an abundance of work in the 1980s on learner modeling for well-defined domains. Those efforts met with considerable success and contributed to the dominance of well-defined domains as a focus for research in AI in education for many years.

Open learner modeling has also proven to be a great strength in learner modeling (Abyaa et al., 2019; Bodily et al., 2018; Bull & Kay, 2007). With the goal of making adaptive learning environments' representations of student competencies inspectable to students, open learner models have emerged as an increasingly attractive family of learner modeling functionalities. While there are no doubt computational and user experience design challenges remaining to be addressed, all indicators suggest that open learner modeling will continue to make great headway going forward.

Weaknesses

Despite major advances that are continuing unabated, learner modeling also suffers from significant weaknesses. Many of these stem from the relatively narrow range of target learner phenomena that have been modeled. Because learner modeling has historically had a strong cognitive focus, cognitive learner modeling capabilities have grown at a steady pace, while other learner modeling capabilities have lagged behind. Affective learner modeling (Hernandez et al., 2010; Yadegaridehkordi et al., 2019) has been the focus of increasing attention, but it is far behind cognitive learner modeling. Although learner affect unequivocally plays a critical role in learning, and even though two decades of work have produced significant advances, affective learner modeling remains considerably weaker than cognitive learner modeling.

While limited work has been done on affective learner modeling, work on metacognitive learner modeling is even more limited. Although a good bit of work has been underway on metacognition and AI in education (Azevedo, 2005; Azevedo et al., 2010) and self-regulated learning (Nietfeld et al., 2014; Sabourin et al., 2013a; Sabourin et al., 2013b; Segedy et al., 2015; Shores et al., 2009; Taub et al., 2016; Taub et al., 2020), we have not established a core set of metacognitive learner modeling functionalities, nor have we created standard representational and inferential frameworks for encoding and reasoning about metacognitive states and abilities, despite the extraordinary importance of metacognition for most learning tasks and contexts, and most learner populations. It should also be noted that younger learner populations pose significant challenges for metacognitive learner modeling, both because metacognition in younger learners is not fully developed and because observing and drawing inferences about metacognitive processes in younger learners is particularly difficult. Further, designing metacognitive learner models for attention and awareness, as well as for reflection (Carpenter et al., 2020) and planning, is not well understood.

Another weakness in current learner modeling is the limited research that has moved beyond learner modeling for individual students. Collaborative learning is highly effective, and collaboration is a twenty-first century competency that is essential for students to acquire (Laal et al., 2012; Smith et al., 1992). However, almost all learner modeling is designed to represent and draw inferences about students learning "solo." While it might seem that classic learner modeling methods could be used for groups of learners, e.g., creating one learner model for each of a pair of learners in a dyad or creating three learner models for each learner in a triad, this approach would fail to capture the dynamics of collaborative learning. Computer-supported collaborative learning (Jeong et al., 2016; O'Malley et al., 2012; Pugh et al., 2021; Sun et al., 2020) introduces the opportunity (and the need) to model student communication, coordination, and group dynamics. For example, game-based collaborative learning (Saleh et al., 2019; Saleh et al., 2022) calls for collaborative learner modeling. However, our understanding of how to model collaboration phenomena is highly underdeveloped compared to conventional learner modeling. At least two types of capabilities are currently missing: modeling collaborative learning phenomena during the course of collaborative learning, and modeling students' collaboration competencies *per se*. Creating learner models that provide these capabilities presupposes learner model designs that are grounded in sociocultural theories of learning (Wang et al., 2019; Danish & Gresalfi, 2018), which is challenging because the vast majority of work on learning modeling has been driven by cognitivist learning theory.

Another weakness of conventional learner modeling is its limitations in supporting learning in ill-defined domains. As noted above, well-defined domains such as mathematics readily lend themselves to straightforward learner modeling approaches, but modeling students learning for ill-defined subject matters poses considerable challenges. For example, how should learner modeling operate in adaptive learning environments for teaching skills such as negotiation, persuasion, public speaking, conflict management, leadership, and social skills more generally?

A final notable weakness of learner modeling is its use of impoverished data streams. Since the early days of research on AI in education, coarse performance data, such as students' responses to multiple choice questions, have been captured because that was all that was available. However, with the emergence of rich data streams spanning video, audio, biometrics, and granular behavior trace data, such as that generated by game-based learning environments (Rowe et al., 2011), it is evident that current learner modeling methods have not caught up with the availability of data produced by current learning environments.

Opportunities

Opportunities abound for learner modeling. As a result of current and future advances in underlying learning environment technologies, as well as a consequence of emerging conceptualizations of learning, learner modeling research can profitably take many directions. Narrative-centered learner modeling presents many opportunities (Lester et al., 2014; Mott et al., 1999). For example, as powerful narrative-centered learning environments emerge, they will generate granular story-driven learning interaction traces. Narrative-centered learner modeling frameworks can then be developed that leverage these rich data to model a wide range of student competencies. Narrative-centered learning environments can also be driven by multi-timescale story-based problem-solving interactions playing out over seconds, minutes, weeks, and perhaps even months. These multi-timescale narrative episodes can provide insight into student learning that spans both cognitive and affective components of learning. Further, the data from these multi-timescale narrative episodes can model both cognitive and affective components of narrative-centered learning.

Advances in pedagogical agents (Johnson & Lester, 2016; Johnson & Lester, 2018; Johnson et al., 2000) are introducing unparalleled opportunities for learner modeling. Pedagogical agents with full spoken language communication capabilities complemented by a broad array of non-verbal communication capabilities (both interpretation and synthesis) will enable a new generation of embodied conversational learner modeling. Learner models will be driven by mentor agents, learning companions, and teachable agents that can engage in rich dialogue, which can then drive robust inference in learner models for learners' knowledge, goals, plans, and preferences. Facilitator agents that interact with both students and teachers can further increase the inferential power of embodied conversational agent learner modeling.

Accelerating improvements in natural language processing create opportunities for new forms of learner modeling. They open possibilities for text-based learner modeling that can draw inferences about students' text-based reflections (Geden et al., 2021), as well as their text-based short answers and essays (Ramalingam et al., 2018; Putra et al., 2021). Integrating evidence provided by analyses of student text will significantly strengthen learner models previously relying on conventional data streams. Text-based conversational dialogue (Min et al., 2016; Min et al., 2019; Wiggins et al., 2019) also introduces new possibilities for learner modeling, and spoken language dialogue-based learner modeling will provide new opportunities for learner models as well, including those utilized in conjunction with embodied conversational agents as described above. These also include new possibilities for affective learner modeling through prosody and sentiment.

Biometrics and expanding sensor capabilities also create new opportunities for affective learner modeling. As biometric sensors become increasingly commoditized and, therefore, increasingly available, they will enable affective learner modeling that is not only more powerful than what we have today, but also delivers these capabilities in a broader range of learning contexts. Emerging biometric sensor technologies also introduce the opportunity to create learner models for psychomotor skills that operate at levels of granularity that were, until very recently, fully unimaginable.

New virtual reality (VR) and augmented reality (AR) technologies create extraordinary possibilities for "full-presence" learner modeling. Integrating adaptive learning technologies with VR and AR creates the opportunity to design learner modeling capabilities that are deeply responsive to learners in immersive environments, which will provide voluminous spatial and temporal data about learners' movement and their interaction with artifacts and other learners. This in turn will contribute to learner models that can accurately reason about learners with considerably richer sources of information than learner models of the past had access to.

Multi-context learner modeling presents significant opportunities as well. Rather than developing learner models for only formal learning contexts (e.g., K-12 school classrooms, Army school houses) or informal learning contexts (e.g., museums), beginning to instrument learners as they move across learning contexts introduces the possibility to create learner models that can leverage evidence about learner competencies in a variety of settings. As a result, one can imagine learner models that are inherently "portable," i.e., they can effectively model students in multiple settings and even draw on evidence from learners interacting in previous settings for transfer to new settings.

Finally, while lifelong learning and lifelong learner modeling have long been considered the ultimate challenge for learner modeling, they also represent the ultimate opportunity. It will soon be technically feasible for learner models to operate at timescales of years. The prospect of lifelong learner modeling of course raises many serious concerns, as noted below, but it also introduces the opportunity to create student-adaptive learning experiences that can draw on an enormous amount of learner experience for unprecedented levels of pedagogical tailoring.

Threats

Rapidly advancing learner modeling capabilities pose significant threats. Many of these threats center on issues of fairness, accountability, and transparency (Gardner et al., 2019; Kizilcec et al., 2020; Paquette et al., 2020). As is widely recognized, learner models are only as good as the data on which they are trained. As a result, training on biased data will produce biased learner models, which can then adapt pedagogy in ways that are far from beneficial and are actually harmful. It will thus be essential to preemptively address learner model bias through learner model training. In a similar vein, it will be important to formulate policy around learner modeling accountability. For example, what organizations and parties are responsible if a learner model were to operate with prejudicial behavior, and how can we formulate policy that most effectively addresses these learner modeling issues before they occur?

The lack of transparency in learner modeling poses a significant threat as well. While learner models that use classic machine learning frameworks are typically transparent, learner modeling frameworks that are based on deep learning are not. As deep learning becomes increasingly powerful, it will likely see rapid uptake in learner modeling, which does not bode well for transparency. In the same ways that other machine learning-based models utilizing deep learning require transparency, such as those in healthcare, finance, and law, so does machine learning-based learner modeling. Further, learner models must serve many stakeholders, including students, teachers, administrators, and parents, placing an even greater burden on model transparency.

Finally, ownership issues in learner models pose significant threats. The question of who owns the data in a learner model will be the subject of increasingly vigorous debate. The learner model represents a particular student's competencies, so it seems the student would own the data. However, school districts may argue that they own the data, and industry will no doubt make similar claims. Questions about where the learner model data resides (or where it should reside) will likely further complicate ownership issues, all of which intersect heavily with the privacy issues noted above.

Overall SWOT Analysis

Table 1 presents a summary of the strengths, weaknesses, opportunities, and threats for learner modeling research.

Table 1. Learner modeling SWOT Analysis.

Strengths	**Weaknesses**
• Robust probabilistic learner models • Increasingly powerful machine learning frameworks (e.g., deep learning) • Learner modeling for well-defined domains • Open learner modeling	• Affective learner modeling • Metacognitive learner modeling • Collaborative/team learner modeling • Learner modeling for ill-defined domains • Impoverished data streams
Opportunities	**Threats**
• Narrative-centered learner modeling • Learner modeling with pedagogical agents • Learner modeling with natural language processing • Learner modeling with biometric and sensor technologies • VR/AR learner modeling • Multi-context learner modeling • Lifelong learner modeling	• Fairness and bias in learner modeling • Learner modeling transparency • Accountability and ownership in learner modeling

Discussion and Recommendations for Future Research

We have reached a critical juncture in the history of learner modeling research. With the advent of deep learning-based models, we have seen rapid advances in the capabilities of learner modeling. Progress in deep learning-based learner modeling frameworks will no doubt continue, with previous successes in well-defined domains being extended to ill-defined domains. We will also see the continued emergence of increasingly sophisticated open learner models, which will have intelligent user interfaces mediating interactions between learners and powerful learner model backends. Despite these advances, it is imperative that learner modeling research address the weaknesses pervading current work. The field needs to address significant deficiencies in affective learner modeling, metacognitive learner modeling, and collaborative/team learner modeling. It must also push on limitations in modeling learners' knowledge components and skills in ill-defined domains and expanding the range of data streams that can inform learner modeling.

The field is now presented with unprecedented opportunities. With the emergence of a new generation of AI-driven narrative-centered learning environments, we can now envision, create, and experiment with narrative-centered learning models. In addition, as embodied conversational agents become increasingly capable of engaging in robust conversational interactions with learners, we can devise pedagogical agent-driven learner models that are deeply informed by verbal and non-verbal communication. More generally, dramatic increases in natural language capabilities will support both speech-based and text-based learner models. The field should also leverage new opportunities introduced by the emergence of VR and AR to fundamentally re-envision how learners interact with learner models. While pursuing these opportunities, it is essential that we address core issues in fairness, accountability, and transparency of learner models, as adaptive instructional systems are only as good as the learner models that drive them.

Recommendations for GIFT Overall

Given the strengths, weaknesses, opportunities, and threats discussed above, we recommend that future work on GIFT address five key areas in learner modeling. First, collaborative learner modeling and team learner modeling will be central to adaptive instructional systems going forward. We can no longer rely on single-learner learner models; we must enable GIFT to have access to team learner models that explicitly represent and draw inferences about team knowledge, skills, and problem-solving strategies. Second, as increasingly powerful narrative-centered learning environments come online, enabling GIFT to have learner models that are "narrative-native" will introduce learner modeling capabilities that operate effectively in scenario-based education and training. Third, GIFT should have robust NLP-driven learner modeling capabilities that support learner modeling from text and speech. With the emergence of pedagogical agents that can engage in rich conversation, which often will take place in narrative-centered learning episodes, GIFT will need to provide spoken-language learner modeling that enables it to draw inferences about learners through language. Fourth, as additional interaction modalities become increasingly common, GIFT will need to support multimodal learner modeling. For example, trainers will likely expect GIFT to model trainee skills based on interaction data from VR-based and AR-based experiences. Finally, GIFT will need to support learner model transparency. While learner models that are opaque will be common as a result of deep learning-based models' growing prominence, explainability in GIFT learner models will become increasingly valuable.

Conclusions

Learner modeling has an extraordinarily bright future. With accelerating advances in AI, learner models will assuredly become more accurate, efficient, and ubiquitous. Performance on machine learning, natural language processing, and computer vision benchmarks will continue their dramatic climbs, which will produce increasingly powerful learner modeling frameworks. With rapid improvements in machine learning for temporal and spatial data and parallel advances in multimodal machine learning, we will see the emergence of learner models with exceptional predictive accuracy for a wide range of learning phenomena, learner populations, and learning contexts.

Soon we will see a rapid succession of new generations of learner models that offer fundamentally new capabilities. These will begin appearing in education and training systems operating in the field, and the data they generate will feed a virtuous cycle of ever more powerful learner models that support an increasingly wide range of adaptive learning. Because of the enormous impact that learner models will have on the effectiveness of adaptive learning environments, continuing successes will fuel the demand for even greater capabilities and introduce even more opportunities. At the same time, the threats noted above will play out on an even larger stage, which will no doubt force innovation in both technology and policy. Today's learner models, whose capabilities far exceed those of years past, will pale in comparison to those of even a few years from now. Together, these developments will promote the creation of learner models that support extraordinarily effective education and training.

Acknowledgements

This research was supported by the U.S. Army Research Laboratory under cooperative agreement #W912CG-19-2-0001. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army.

References

Abyaa, A., Khalidi Idrissi, M., & Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. *Educational Technology Research and Development*, *67*(5), 1105–1143. https://doi.org/10.1007/s11423-018-09644-1

Azevedo, R., Moos, D. C., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist, 45*(4), 210–223. https://doi.org/10.1080/00461520.2010.515934

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199–209.

Biswas, G., Rajendran, R., Mohammed, N., Goldberg, B. S., Sottilare, R. A., Brawner, K., & Hoffman, M. (2019). Multilevel learner modeling in training environments for complex decision making. *IEEE Transactions on Learning Technologies*, *13*(1), 172–185. https://doi.org/10.1109/TLT.2019.2923352

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: a systematic review. In Shum, S. B., Ferguson, R., Merceron, A., & Ochoa, X. (Eds.), *Proceedings of the eighth international conference on learning analytics and knowledge* (pp. 41–50). The Association for Computing Machinery. https://doi.org/10.1145/3170358.3170409

Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI:() Open learner modelling framework. *International Journal of Artificial Intelligence in Education*, *17*(2), 89–120. https://doi.org/10.5555/1435369.1435371

Carpenter, D., Geden, M., Rowe, J., Azevedo, R., Lester, J. (2020). Automated analysis of middle school students' written reflections during game-based learning. In Bittencourt, I. I., Cukurova, M., Muldner, K., Luckin, R. & Millán, E. (Eds.), *Proceedings of the 21st international conference on artificial intelligence in education* (pp. 67–78). Springer, Cham. https://doi.org/10.1007/978-3-030-52237-7_6

Carr, B., & Goldstein, I. P. (1977). Overlays: A theory of modelling for computer aided instruction. *Massachusetts Institute of Technology Cambridge Artificial Intelligence Lab.*

Chrysafiadi, K., & Virvou, M. (2013, September). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729. https://doi.org/10.1016/j.eswa.2013.02.007

Danish, J. A., & Gresalfi, M. (2018). Cognitive and sociocultural perspective on learning: Tensions and synergy in the learning sciences. In Fischer, F., Hmelo-Silver, C. E., Goldman, S. R., & Reimann, P. (Eds.), *International handbook of the learning sciences*, (pp. 34–43). Taylor & Francis.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In Ferguson, R., Hoppe, U., & Brooks, C. (Eds.), *Proceedings of the ninth international conference on learning analytics & knowledge* (pp. 225–234). https://doi.org/10.1145/3303772.3303791

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive Student Modeling in Game-Based Learning Environments with Word Embedding Representations of Reflection. *International Journal of Artificial Intelligence in Education, 31*(1), 1–23.

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, *12*(3), 31–54.

Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2021). Multimodal Multi-Task Stealth Assessment for Reflection-Enriched Game-Based Learning. In Di Mitri, D., Martínez-Maldonado, R., Santos, O. C., Schneider, J.,
Mat Sanusi, K. A., Cukurova, M., Spikol, D., Molenaar, I., Giannakos, M., Klemke, R., & Azevedo, R. (Eds.), *Proceedings of the first international workshop on multimodal artificial intelligence in education (MAIED 2021) at the 22nd international conference on artificial intelligence in education (AIED 2021)* (pp. 93–102). CEUR Workshop Proceedings.

Hernández, Y., Sucar, L. E., & Arroyo-Figueroa, G. (2010). Evaluating an affective student model for intelligent learning environments. In A. Kuri-Morales & G. R. Simari (Eds.), *Advances in artificial intelligence –*

IBERAMIA 2010 (pp. 473–482). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16952-6_48

Jeong, H., & Hmelo-Silver, C. E. (2016). Seven affordances of computer-supported collaborative learning: How to support collaborative learning? How can technologies help? *Educational Psychologist, 51*(2), 247–265. https://doi.org/10.1080/00461520.2016.1158654

Johnson, W. L., & Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial Intelligence in Education, 26*(1), 25–36. https://doi.org/10.1007/s40593-015-0065-9

Johnson, W. L., & Lester, J. C. (2018). Pedagogical agents: back to the future. *AI Magazine, 39*(2), 33–44. https://doi.org/10.1609/aimag.v39i2.2793

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11*(1), 47–78.

Kizilcec, R. F., & Lee, H. (2020). *Algorithmic fairness in education*. arXiv. https://doi.org/10.48550/arXiv.2007.05443

Laal, M., & Ghodsi, S. M. (2012). Benefits of collaborative learning. *Procedia - Social and Behavioral Sciences, 31*, 486–490. https://doi.org/10.1016/j.sbspro.2011.12.091

Lester, J. C., Spires, H. A., Nietfeld, J. L., Minogue, J., Mott, B. W., & Lobene, E. V. (2014). Designing game-based learning environments for elementary science education: A narrative-centered learning perspective. *Information Sciences, 264*, 4–18. https://doi.org/10.1016/j.ins.2013.09.005

Min, W., Wiggins, J. B., Pezzullo, L. G., Vail, A. K., Boyer, K. E., Mott, B. W., Frankosky, M. H., Wiebe, E. N., & Lester, J. C. (2016). Predicting dialogue acts for intelligent virtual agents with multimodal student interaction data. In Barnes, T., Chi, M., & Feng, M. (Eds.), *Proceedings of the ninth international conference on educational data mining,* (pp. 454–459).

Min, W., Park, K., Wiggins, J. B., Mott, B., Wiebe, E., Boyer, K. E., & Lester, J. (2019). Predicting dialogue breakdown in conversational pedagogical agents with multimodal LSTMs. In Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B. & Luckin, R. (Eds.), *Proceedings of the 20th international conference on artificial intelligence in education*, (pp. 195-200). Springer. https://doi.org/10.1007/978-3-030-23207-8_37

Mott, B. W., Callaway, C. B., Zettlemoyer, L. S., Lee, S. Y., & Lester, J. C. (1999). Towards narrative-centered learning environments. In Mateas, M. & Sengers, P. (Eds.), *Proceedings of the 1999 AAAI fall symposium on narrative intelligence* (pp. 78–82). The AAAI Press.

Nietfeld, J. L., Shores, L. R., & Hoffmann, K. F. (2014). Self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology, 106*(4), 961–973. https://doi.org/10.1037/a0037116.

O'Malley, C. (Ed.). (2012). Computer supported collaborative learning (Vol. 128). Springer Science & Business Media.

Owen, V. E., Roy, M. H., Thai, K. P., Burnett, V., Jacobs, D., Keylor, E., & Baker, R. S. (2019). Detecting Wheel-Spinning and Productive Persistence in Educational Games. In Lynch, C. F., Merceron, A., Desmarais, M., & Nkambou, R. (Eds.), *Proceedings of the 12th international conference on educational data mining*, (pp. 378-383).

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining, 12*(3), 1–30.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3), 313–350. https://doi.org/10.1007/s11257-017-9193-2

Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., & D'Mello, S. K. (2021). Say what? Automatic modeling of collaborative problem solving skills from student speech in the wild. In Hsiao, I., Sahebi, S., Bouchet, F. & Vie, J. (Eds.), *Proceedings of the 14th international conference on educational data mining*, (pp. 55-67).

Putra, J. W. G., Teufel, S., & Tokunaga, T. (2021, April). Parsing argumentative structure in english-as-foreign-language essays. In Burstein, J. Horbach, A., Laarman-Quante, R., Leacock, C., Madnani, N. Pilàn, I., Yannakoudakis, H., & Zesch, T. (Eds.), *Proceedings of the 16th workshop on innovative use of NLP for building educational applications* (pp. 97–109). Association for Computational Linguistics.

Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018, April). Automated essay grading using machine learning algorithm. *Journal of Physics: Conference Series*, *1000*(1), 012030. https://doi.org/10.1088/1742-6596/1000/1/012030

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education, 21*(1–2), 115–133. https://doi.org/10.3233/JAI-2011-019

Sabourin, J. L., Shores, L. R., Mott, B. W., & Lester, J. C. (2013a). Understanding and predicting student self-regulated learning strategies in game-based learning environments. *International Journal of Artificial Intelligence in Education, 23*(1), 94–114. https://doi.org/10.1007/s40593-013-0004-6

Sabourin, J., Mott, B., & Lester, J. (2013b). Utilizing dynamic Bayes nets to improve early prediction models of self-regulated learning. In Carberry, S., Weibelzahl, S., Micarelli, A., & Semeraro, G. (Eds.) *Proceedings of the 21st international conference on user modeling, adaptation, and personalization* (pp. 228–241). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38844-6_19

Saleh, A., Chen, Y., Rehmat, A., Housh, K., Hmelo-Silver, C., Glazewski, K., & Lester, J. (2019). Assessing collaborative problem solving in the context of a game-based learning environment. In Lund, K., Niccolai, G. P., Lavoué, E., Hmelo-Silver, C., Gweon, G., & Baker, M. (Eds.), *Proceedings of the 13th international conference on computer-supported collaborative learning, volume 2* (pp. 893–894). International Society of the Learning Sciences.

Saleh, A., Phillips, T. M., Hmelo-Silver, C. E., Glazewski, K. D., Mott, B. W., & Lester, J. C. (2022). A learning analytics approach towards understanding collaborative inquiry in a problem-based learning environment. *British Journal of Educational Technology.* https://doi.org/10.1111/bjet.13198

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2015). Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics, 2*(1), 13–48.

Shores, L. R., Robison, J. L., Rowe, J. P., Hoffman, K. L., & Lester, J. C. (2009). Narrative-centered learning environments: A self-regulated learning perspective. In Pirrone, R., Azevedo, R., & Biswas, G. (Eds.), *Cognitive and metacognitive educational systems: Papers from the AAAI fall symposium (FS-09-02)* (pp. 87-92). The AAAI Press.

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

Smith, B. L., & MacGregor, J. T. (1992). What is collaborative learning? In Goodsell, A. S., Maher, M. R., Tinto, V., Smith, B. L. & MacGregor, J. T. (Eds.), *Collaborative learning: A sourcebook for higher education*, (pp. 10-30). National Center on Postsecondary Teaching, Learning and Assessment.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education, 143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2016). Using multi-level modeling with eye-tracking data to predict metacognitive monitoring and self-regulated learning with Crystal Island. In Micarelli, A., Stamper, J., & Panourgia, K. (Eds.), *Proceedings of the 13th international conference on intelligent tutoring systems* (pp. 240–246). Springer, Cham. https://doi.org/10.1007/978-3-319-39583-8_24

Taub, M., Sawyer, R., Lester, J., & Azevedo, R. (2020). The impact of contextualized emotions on self-regulated learning and scientific reasoning during learning with a game-based learning environment. *International Journal of Artificial Intelligence in Education, 30*(1), 97–120. https://doi.org/10.1007/s40593-019-00191-1

VanLehn, K. (1988). Student modeling. *Foundations of Intelligent Tutoring Systems*, *55*, 78.

Wang, M. T., Degol, J. L., & Henry, D. A. (2019). An integrative development-in-sociocultural-context model for children's engagement in learning. American Psychologist, 74(9), 1086–1102. https://doi.org/10.1037/amp0000522.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge*. Morgan Kaufmann.

Wiggins, J. B., Kulkarni, M., Min, W., Boyer, K. E., Mott, B., Wiebe, E., & Lester, J. (2019). Take the initiative: Mixed initiative dialogue policies for pedagogical agents in game-based learning environments. In Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B. & Luckin, R. (Eds.), *Proceedings of the 20th international conference on artificial intelligence in education* (pp. 314–318). Springer, Cham.

Yadegaridehkordi, E., Noor, N. F. B. M., Ayub, M. N. B., Affal, H. B., & Hussin, N. B. (2019). Affective computing in education: A systematic review and future research. *Computers & Education, 142*, 103649.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In Lane, H. C., Yacef, K., Mostow, J., & Pavlik, P. (Eds.), *Proceedings of the 16th international conference on artificial intelligence in education* (pp. 171–180). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39112-5_18

Chapter 4 - Instructional Strategies in Intelligent Tutoring Systems SWOT Analysis

Jong W. Kim[1], Steve Ritter[2], Michael Krusmark[1], and Tiffany S. Jastrzembski[3]
CAE USA[1]; Carnegie Learning[2]; US Air Force Research Laboratory (AFRL)[3]

Introduction

Instructional strategies can be thought of as a two-layer process including an outer loop, and an inner loop in a computer-guided training system. In the outer loop, strategies can be applied to select learning activities that are generally specified in a goal-oriented learner model. That is, the outer loop ideally can support the different activities for different learners, in recognition that knowledge builds on prior knowledge. During this outer loop process, the inner loop process supports achieving the learning goals by providing assessment and feedback and by supporting individual student strategies to complete the activity. Knowledge component modeling can support both the inner and outer loops (Goldin, et al., 2016). Within the inner loop, steps taken in the activity are associated with knowledge components (KCs), allowing the system to build up a profile of the components on which students are proficient and those that still need more practice. When the activity is completed, outer loop processes can use this knowledge component profile to select an appropriate next activity for the student. This two-layer process can be implemented as an intelligent tutoring system (ITS) to help the learner to complete the learning activities and objectives. The instructional strategies can be further reinforced by using the Artificial Intelligence (AI)/Cognitive Science and machine learning (ML) techniques to improve learning and retention of knowledge and skills—e.g., spaced practice trials would slow learning but increase retention (Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018). In this chapter, based on the SWOT (Strengths, Weakness, Opportunities, and Threats) analysis, we describe research-supported understandings with regard to learning data analytics and predictive tools for advancing instructional strategies.

With inner and outer loop strategies, improving instructional strategies in an ITS would require two major capabilities: (a) ability to deal with multi-skills learning, and (b) ability to advance prescriptive personalized learning for the individual learner. Instructional strategies with personalization and multi-skills acquisition would need symbiotic interplay of the outer and inner loop processes. They can be bolstered by predictive analytics on learning and decay – i.e., when the learner needs an additional practice trial for reinforced retention. The current technologies of computational cognitive models with the support of AI/ML techniques have started to address this research question (e.g., Sense, Wood et al., 2021), and to support such predictions to an extent. Based on these points, we will provide instructional strategies in terms of SWOT analytics for further research and integration with an ITS (e.g., GIFT, or other relevant platforms including AutoTutor, OpenTutor, and D2P).

It can be helpful to think of three perspectives on knowledge components: cognitive, educational and analytic. From a *cognitive perspective* (see Koedinger et al., 2012), KCs reflect the underlying mechanism that the brain uses to solve problems. KCs represent mental process as well at the 10 s (seconds) unit task level in Newell's time band (Newell, 1990); unit tasks usually last 10 s. Furthermore, they could be ideally represented in cognitive constructs – i.e., production rules with a set of declarative knowledge in a cognitive architecture, ACT-R (Anderson, 2002; Anderson et al., 2004). While, in theory, the complete set of KCs required to complete a task could be represented in a domain model, in practice, there are far too many, so systems tend to represent a small subset of the cognitive processing that is actually required to complete a task.

The *educational perspective* on KCs focuses on the educational intent of an instructional system. Such systems often have external educational objectives (also called competencies or standards in different educational contexts). KCs represent a finer-grained view of the educational objectives of a system. For example, one of the Common Core State Standards for Mathematics used in several US states is "8.ee.7b: Solve linear equations with rational number coefficients, including equations whose solutions require expanding expressions using the distributive property and collecting like terms." This standard expresses the educational goal, but it is too broad to provide instruction because, for students to master this standard requires the student to learn and practice many underlying KCs. For example, in a system like MATHia (Ritter et al., 2007), solving equations of the form "$ax = b$" involves KCs representing the ability to solve such equations when a is a positive integer, when a is a negative integer, when a is -1 (a special case because mathematical notation writes -1x as -x) and several other cases, in addition to the KCs underlying more complex linear equations. These KCs represent difficulty factors (Baker et al., 2007) that make some equations harder for some students to solve than others and also help to divide the space of activities (equations to solve) in a way that allows the outer loop to sample it much more efficiently than picking problems from the space of problems represented by the standard 8.ee.7b.

There are many possible partitions of a standard like 8.ee.7b that could be used to guide instruction. The *analytic perspective* on KCs helps to guide this partitioning in a way that is most educationally efficient. Building on the cognitive perspective, we treat the KCs as parameters in a model that predicts student learning and performance (Cen et al., 2006; Goldin et al., 2016). If, as in the cognitive perspective, KCs are the things that get learned (and get better with practice), then we should see KCs improve along a power law of learning, which is that the time to complete a task speeds up with practice according to a power function. If we see deviations from power law of learning (and, particularly, if we see drops in learning for particular activity types, then we know that the modeled KCs do not correspond to the actual KCs that students are using to solve problems.) This perspective explains why we treat equations with positive coefficients differently from equations with negative coefficients. There is no logical or mathematical reason to do so, but, empirically, the data show that students at this level in their education who are able to solve $ax = b$ equations where a is a positive integer may not be able to solve such equations when a is a negative integer. Data-driven discovery of new knowledge components thus leads to improvements in the efficiency of educational systems (Liu & Koedinger, 2017).

Combining these perspectives, we can see that the function of KCs is to generalize terms for describing concepts, facts, cognition, and knowledge in a way that provides guidance to the outer loop in activity selection and also supports a foundation for data-enabled assessment and improvement in learning. The KC modeling approach provides a practical connection between learning science and education.

However, assessment of learner performance is somewhat limited – assessment in ITSs provide only a snapshot of current capabilities of the learner rather than a prediction for the future (Pavlik Jr et al., 2017). It is necessary to understand what an assessment might imply for long-term proficiency for improved predictive readiness. The two approaches of multi-skills learning, and predictions on personalized learning and decay using *Predictive Performance Equations* (PPEs) can help us to instantiate such technical gaps of instructional strategies (Gluck et al., 2019). Using PPE, we can better track KCs that can be interpreted by learning performance, prediction on learning, and retention. PPEs can address the spacing effect to predict when to retrain (e.g., Gluck, et al., 2019; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018). We continue improving PPEs to capture both different learning rates and decay rates in terms of the process of declarative to procedural learning and retention. In terms of the three stages of learning and retention, the learner goes through from the declarative to the procedural stage (which is summarized in a cognitive theory, and implemented in a training architecture, D2P). A large complex task can be decomposed into multiple subtasks and subskills. We need to be aware that they might be learned differently and be forgotten differently. These attempts can help to achieve our mutual goal of improving personalization, and predicting optimal training schedules for longer retention.

When it comes to tracing KCs, as a formative assessment technique, learning (and forgetting) curves can provide important insights of how to assess performance and define adaptivity levels. In general, task completion time follows a power law of learning (and forgetting) representing a speed-up effect. Assessment of performance can be ranging from milliseconds to years; spanning of seven orders of magnitude (Anderson, 2002). This is useful in designing an adaptive instructional system since a large task can be meaningfully decomposed.

One of known pitfalls of learning (forgetting) curves is that a larger domain model or a large student sample size is likely to exhibit a better fit than a smaller one, even if the system does not teach the students any better (Martin et al., 2011). For example, a larger task can be decomposed, but subtasks would be learned differently (Kim & Ritter, 2016). Thus, a simple comparison of a learning (forgetting) curve in a large task seems not sufficient enough to support personalized instructional strategies. Furthermore, a near-term assessment by comparing learning (forgetting) curves would not be related to the long-term stability of learning (e.g., Schmidt & Bjork, 1992).

SWOT Analysis

We need reliable predictions on learning, and particularly decay as well to advance inner and outer loop behavior of an ITS. Scientists demonstrate that learning happens in stages, and the process of learning and forgetting can be represented by declarative and procedural knowledge in a cognitive model (Anderson, 1993). One of the clearly sharable goals by domains in Cognitive Science and Education is to provide improved learning and longer retention of acquired knowledge and skills by the stages of learning and retention (see Kim & Ritter, 2015) . To achieve the goal, more personalized instructional strategies would be necessary because one size does not fit all, and instructional schedules should be optimized accordingly. When do we need massed or spaced practice? Do we really know what is learned and forgotten in an item level detail? When do we need to retrain? All these fundamental research questions are related to improving instructional strategies. We analyze this issue by taking a SWOT analysis approach to succinctly summarize technical challenges and scientific directions.

Strengths

Knowledge component modeling is useful to examine performance changes and guide data-driven improvement (Xiangen Hu, Instructional Strategies PADLET, 2021). A learner model with KCs can specify how the learner would acquire knowledge and skills (probably multi-skills) in a task through the stages of learning (e.g., from declarative to procedural stages). The learner model can generalize the terms for describing pieces of cognition or knowledge including production rules, facts, principles, concepts, and schema (Koedinger et al., 2012). The model can be implemented as a rule-based cognitive model that can track student learning and performance in real time. A production rule-based model can help in thinking about what knowledge may be needed to perform a particular task, how that knowledge might be decomposed to capture what the learner would do, and how widely specific knowledge components will transfer (Aleven & Koedinger, 2013).

PPE, Predictive Performance Equation (Gluck et al., 2019), has been developed to pursue the goal to trace and predict decay for future performance – i.e., predictive information about when to relearn, seeking optimized training schedules. Based on the KC-based model in an ITS, we can trace the process of learning and forgetting to prescribe learning schedules for longer retention; data tagged by each KC over time can be collected from the laboratory or the field.

A larger task can be meaningfully and functionally decomposed to smaller unit tasks (Lee & Anderson, 2001). Similarly, multi-skills can be decomposed for meaningful analytics that allow us to predict student performance. Multi-skills can be decomposed into measurable units of knowledge and skills that can represented as KCs. Integrating a KC model in an ITS can support advanced learning analytics to predict learning and decay.

Table 1. A summary of strengths.

Knowledge component modeling is useful to measure performance changes in an ITS.
Knowledge component models help an ITS to be more data-driven.
GIFT can incorporate a knowledge component model (e.g., PPE) to provide adaptive training schedules.

Weaknesses

In cognitive science, scientists generally investigate performance changes (declarative to procedural) in milliseconds, but educational outcomes could be months and years. That is, assessment of performance can be ranging from milliseconds to years; there are spanning orders in magnitude. Tracing KCs needs to be specified in detail by addressing the spanning orders, so that they can be computationally implemented in the inner and outer loops of ITSs. This scalability issue can further bring: (a) time scale difference from inner and outer loops, (b) differences in scales of hierarchical multi-skills levels, and (c) individual variations for learning and decay. To successfully utilize PPE in ITSs, we need adequate data fidelity, based on objective performance metrics, and data from repeated measures to assess learning and decay. AFRL has conducted extensive research on fatigue (e.g., sleep deprivation), and skill acquisition. The lab findings need to be generalized to the field, addressing scalability issues. In addition, computational costs for simulation and prediction can be weaknesses when knowledge and skills with KCs and the number of the learners in the wild are taken into consideration.

KCs are best used for educational objects that constitute domain knowledge and require proceduralization. They provide poor support for learning objectives where practice is relatively unimportant (for example, a short exposure to a concept like absolute value might be sufficient and thus more efficient than a serious of tasks related to the concept), and they may be less useful for tracking more general strategies or approaches to learning. For example, in addition to content standards focused on things like solving linear equations, the Common Core State Standards for Mathematics include “Standards for Mathematical Practice” like “make sense of problems and persevere in solving them”. While it may be possible to model learning of such objectives, it is an open question as to whether they have the same psychological reality as domain knowledge from the cognitive perspective on KCs. Thus, such objectives might not follow the same kind of power law of learning. At a practical level, mapping performance on a task to a discrete instance of “persevering,” for example, poses challenges as well.

Table 2. A summary of weaknesses.

There is a gap between outcomes in cognitive level analytics and educational outcomes.
Knowledge components might sometimes not be able to fully support the learning objectives.

Opportunities

In general, we should consider the ITS to be part of a modern learning and training environment. This allows us to focus the ITS on what it does best, which is learner work that can be well modeled with KCs, as well as support for instructors. One advantage of the KC approach is that KCs are explainable, so the ITS can produce a learner model that is understandable to both the student and the instructor, and the instructor can author the learning materials for the learner that responds to student failure to master particular KCs. Furthermore, integrating the capability of predicting decay (e.g., using PPE) into the ITS can be a greater opportunity to provide advanced information for the learners and the instructors. There are no training systems that the authors are aware of that can provide varying learning (decay) rates by subtasks and by individual learner. These are the important measures that can enable personalized instructional strategies. The aforementioned perspectives (cognitive, educational, and analytic) let us see significant opportunities for collaboration—i.e., it would be possible to deliver optimized, and cognitively plausible retraining schedules. In addition, we can provide content selection for retraining with consideration of multi-skills learning and decay, achieving enhanced retention.

Table 3. A summary of opportunities.

An ITS can support training schedules based on predictive analytics on learning and forgetting.

Threats

We observe that many Department of Defense (DoD) systems are not equipped to track objective performance in a digitized and automated fashion. Sometimes aggregate measures of a student's language learning are available, but granularity of measures at the item level would be required to improve personalized training. In addition, many DoD systems are not designed to track learning over time (not allowing for repeated measures). In addition, it has been acknowledged that the learning data from the field would engender some poor fit. Bayesian models can be useful, but there might be expensive computational cost to update each individual learner's posterior distributions by each KC.

Table 4. A summary of threats.

A support of larger defense systems can be critical.

Supporting Research

Humans appear to be remarkably good at learning, but they sometimes tend to practice what they do know rather than what they do not know (Atkinson, 1972). Similarly, when the individual learner learns knowledge and skills (e.g., from a golf putting task to a calculus problem solving), the learner needs to know what they are not good at. There is a consensus theory of stages of learning that has been the foundation for a number of tutoring systems (e.g., Anderson et al., 1990; Anderson et al., 1985; Corbett & Anderson, 1995; Ritter et al., 2013; Ritter et al., 2007).

The theoretical account of the learning behavior in the three stages (e.g., Anderson, 1982; Fitts, 1964; VanLehn, 1996) can provide us with important insights and mechanisms to represent forgetting (Kim & Ritter, 2015). It is reasonable to hypothesize that knowledge is forgotten in each stage. In the first stage, mostly declarative knowledge would be degraded. In the second stage, both declarative (e.g., facts consisting of chunks) and procedural knowledge (e.g., production rules to represent steps and sequences) would be degraded. In the third stage, similar to the second stage, both declarative and procedural forms of knowledge would be degraded. These three stages would be continuous. However, for the clarity of theoretical explanation, we describe each stage distinctively in this section. Later, we will introduce how we can put the distinctive three stages together to better represent forgetting.

A method for skill assessment is used to identify what skills (or subskills) individuals are good at or are not good at. The results of an assessment can then be visualized as a learning (and forgetting) curve, which is typically represented as a power function (Newell & Rosenbloom, 1981). A learning curve would be used for a formative and a summative study to improve adaptive instructional systems (Martin et al., 2011). A summative assessment usually happens after the learner has finished being taught about a subject (e.g., a final exam at the end of a semester or at the end of a unit task). In the meanwhile, a formative assessment happens while a student is being taught about a subject, rather than at the end of year or unit of a work, in order to check their progress.

However, this theoretical understanding is not sufficient to analyze and predict the components of knowledge and skills training. Assessments of each knowledge and skill component would be necessary for improved instructional strategies for personalization. It is necessary to identify the skill level of those three components in terms of these learning stages, and that can provide suggestions of dynamic scaffolding and adaptive instructions. The ITS can be an important tool that can reify the aforementioned theoretical perspectives. Based on the identified knowledge and skill components, an improved ITS could tag the knowledge and skill components that need more practice. For example, if the learner fails to do a task, the learner model implemented in the ITS can automatically identify the less practiced skill from all of the knowledge and skill components. Learner models based on Bayesian hidden markov style knowledge tracing (e.g., Baker, Corbett, & Aleven, 2008; Yudelson et al., 2013) and cognitive model based knowledge tracing (Jastrzembski et al., 2006) have been widely investigated to examine the efficacy of ITSs, identifying skill components by formative assessments and automatically tagging them if they are under-practiced for an adaptive instructional strategy. These efforts seek to suggest dynamic scaffolding and adaptive instructions based on the tagged skill components. An intelligent tutoring framework can provide a type of instructional materials (video or text formats).

As mentioned earlier, individuals appear to practice what they do know rather than what they do not know (Atkinson, 1972), suggesting that personalized guidance on learning is necessary to achieve improved learning of multi-level complex skill components. In Airforce Research Laboratory (AFRL), PPE has been developed, which was initially based on cognitive model-driven knowledge tracing. PPE is a multiplicative equation that both contains a learning term and forgetting term.

PPE is originally inspired by cognitive theories of learning and memory, including *General Performance Equation*, GPE (Anderson & Schunn, 2000). It did successfully account for how the amount of study and elapsed time would affect retention. PPE goes beyond the GPE in that it represents how the temporal distribution of practice affects retention as well. PPE accounts for the spacing effect well. The spacing effect is that separating practice repetitions by a delay slows learning but enhances retention. PPE predicts decay, and the spacing effect, which is used to provide refresher training. Walsh, Gluck, Gunzelmann, Jastrzembski, Krusmark, Myung et al. (2018) compared different models on predictions of the spacing effect, showing PPE provides a reliable prediction on decay with regard to massed versus spaced practice.

Currently, the original PPE has been being extensively tested against learning and training data from the field (e.g., a calendar-based CPR training for certification) and integrated with Machine Learning (ML) models to better account for environmental variances (e.g., noisy data on forgetting and data scarcity for a specific longitudinal condition)(Sense, Collins et al., 2021). The spacing effect is known as one of the most scientifically demonstrated learning principles, but it has not been successfully applied to ITSs. In addition, learning materials that are learned in a spaced manner can be relearned quickly; this is called spacing-accelerated relearning (Jastrzembski, Walsh et al., 2018). PPE is the most mature and robust model that can trace knowledge and skill components, and predict decay. Thus, PPE based on theories of cognitive learning and forgetting is the one candidate computational model that can enable personalized adaptive instructions for the ITS.

Discussion and Recommendations for Future Research

Integrating PPE with an ITS (e.g., The Generalized Intelligent Framework for Tutoring; GIFT) will enhance personalization and decay prediction capability. This will strengthen the operational functionality that drives GIFT, such as adaptive course flow. It also helps to identify and assess the learner state of acquiring knowledge and skills and to prescribe learning and relearning schedules in a spaced or massed fashion for longer term stability of knowledge and skills.

A machine learning (ML) model is not able to find structure without sufficient data. By using cognitive models, it is possible to inform the ML model of structure with data – an ensemble modeling approach of using a cognitive model and ML models. PPE, as a computational cognitive model, seeks to take the best advantages of using ML models. There have been gaps between AI/cognitive science and educational outcomes. People have been trying for decades to mind this gap. How can we make better progress?

As one attempt to close the gap, the ensemble modeling approach can address unexplained variance and uncertainty between those two domains. PPE is deterministic with parameters, and is inspired by a cognitive model. Uncertainty and unexplained variances would exist when we utilize PPE to predict the data from the field. We deal with massive field data (e.g., CPR training data from the field) by using predictive models including ML and statistical/probabilistic models. A statistical/probabilistic model (e.g., a Bayesian hierarchical model) is used to infer parameters from the field data with consideration of error and variance as an inverse model. We solve inverse problems by inferring the values of model parameters that are consistent with the field data iteratively in an attempt to reduce the gap.

Robert Sottilare (personal communication, 2021) mentioned how integration of decay in a predictive model might be represented and standardized across various types of tasks. Starting from a cognitive architecture, ACT-R supports cognitively plausible mechanisms and structures for both human learning and forgetting. Adopting the best structure would be useful to represent decay across various types of tasks. But there are limitations. ML models learn well to predict decay for a specific task, but it would fail without sufficient data to train the ML model. If we give a cognitively plausible structure to a ML model, we might be able to address some of the problems in a similar manner.

Recommendations for GIFT Overall

Based on the understanding of the aforementioned instructional strategies, we can provide design recommendations for GIFT and future ITSs. It is necessary to place the ITS as a central position in the learning process, but other non-ITS learning activities including human-led instructions should also be available for the learner all the time. The ITS can be reinforced by building and supporting data-driven improvements in learning environments. Finally, the ITS should be supportive in providing the learners with recommendations of learning contents and personalized optimal schedules for learning (and relearning).

Conclusions

Personalized instructional strategies are intended to help the learner to acquire KCs, to practice them, and to achieve expertise through the progression of stages of a learning curve, which has shown to impact learning in various task domains including procedural troubleshooting tasks, mathematics, physics problem-solving, etc. Personalized instructional strategies are limited to certain learning environments. They are challenged to address the requirements from multiskills learning by varying individuals.

To achieve this, we need an improved method of assessing forgetting for long-term proficiency with different types of knowledge. Our discussion asserts the necessity of an improved framework for assessment and its interpretation when it comes to forgetting. Particularly, it is necessary to provide an improved framework to assess stability and transferability of the acquired knowledge and skills in an unannounced and unobtrusive way. This problem can be approached by using and extending a computer-based tutoring system.

GIFT can support learning and assessment of the knowledge types discussed earlier in this chapter (e.g., simple recognition, cued recall, transfer of knowledge). In GIFT, a hierarchy of concepts, which is implemented in the Domain Module, can be also expanded to deal with the microgenetics of knowledge and skills (e.g., tasks and subtasks, skills and subskills, or movements or submovements). A microgenetic approach to assess forgetting in a computer-based tutoring system will help us to better identify the learner state and support improved knowledge and skill proficiency.

A limitation to note here is that the GIFT modules are only able to assess the current learner state; it is unable to predict the future learner state. With regard to the changing forgetting rates, GIFT is currently incapable of using predictions from a computational model. However, if GIFT were enhanced to predict the rate of forgetting, it could be used to determine and support strategies that are necessary for acquisition of robust knowledge and skills. Therefore, there may be merit to addressing forgetting rates in the developing GIFT system, to better support learning and assessment capabilities, which will help the system to identify better ways to achieve long-term proficiency. Integrating a predictive tool into an ITS is necessary.

References

Aleven, V., & Koedinger, K. R. (2013). Knowledge component (KC) approaches to learner modeling. In *Design Recommendations for Intelligent Tutoring Systems - Learner Modeling* (Vol. 1, pp. 165-182). Orlando, FL: U.S. Army Research Laboratory.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369-406.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science, 26*, 85-112.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111*(4), 1036-1060.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence, 42*, 7-49.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science, 228*(4658), 456-462.

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in Instructional Psychology: Educational Design and Cognitive Science* (Vol. 5, pp. 1-34). Mahwah, NJ: Lawrence Erbaum Associates.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96*(1), 124-129.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2007). The difficulty factors approach to the design of lessons in intelligent tutor curricula. *International Journal of Artificial Intelligence in Education, 17*(4), 341-369.

Baker, R. S. J., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 406-415). Montreal, Canada: Springer.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis: A general method for cognitive model evaluation and improvement. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 164-175). Berlin, Germany: Springer-Verlag.

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

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243-285). New York: Academic Press.

Gluck, K. A., Jastrzembski, T. S., & Krusmark, M. A. (2019). Prospective comments on performance prediction for aviation psychology. In *Improving Aviation Performance through Applying Engineering Psychology* (pp. 79-98): CRC Press.

Goldin, I., Pavlik Jr., P. I., & Ritter, S. (2016). Discovering domain models in learning curve data. In R. Sottilare, A. Grasser, X. Hu, A. Olney, B. Nye & A. Sinatra (Eds.), *Design recommendations for intelligent tutoring - Domain modeling* (Vol. 4, pp. 115-126). Orlando, FL: US Army Research Laboratory.

Hu, X. (2021). *Instructional Strategies Padlet Comment*. Retrieved from https://padlet.com/xiangenhu/qq4hcewdjwiuke1m

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference* (pp. 1498-1508). Orlando, FL: National Training Systems Association.

Jastrzembski, T. S., Walsh, M. M., Krusmark, M. A., Gluck, K., & Gunzelmann, G. (2018). Personalized learning in the wild: Accounting for effects of spacing, retention, and relearning through use of a cognitive model. In *Proceedings of the Applied Human Factors & Ergonomics Annual Meeting*. Orlando, FL.

Kim, J. W., & Ritter, F. E. (2015). Learning, forgetting, and relearning for keystroke- and mouse-driven tasks: Relearning is important. *Human-Computer Interaction, 30*(1), 1-33.

Kim, J. W., & Ritter, F. E. (2016). Microgenetic analysis of learning a task: Its implications to cognitive modeling. In F. E. Ritter & D. Reitter (Eds.), *Proceedings of the 14th International Conference on Cognitive Modeling* (pp. 21-26). University Park, PA: Penn State.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge‐Learning‐Instruction framework: Bridging the science‐practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757-798.

Lee, F. J., & Anderson, J. R. (2001). Does learning a complex task have to be complex? A study in learning decomposition. *Cognitive Psychology, 42*(3), 267-316.

Liu, R., & Koedinger, K. R. (2017). Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining, 9*(1), 25-41.

Martin, B., Mitrovic, A., Koedinger, K. R., & Mathan, S. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction, 21*(3), 249-283.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Lawrence Erlbaum.

Pavlik Jr, P. I., Maass, J. K., & Kim, J. W. (2017). Assessment of forgetting. In R. Sottilare, A. Graesser, X. Hu & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems: Volume 5 - Assessment methods* (Vol. 5, pp. 203-208): U.S. Army Research Laboratory.

Ritter, F. E., Yeh, K.-C., Cohen, M. A., Weyhrauch, P., Kim, J. W., & Hobbs, J. N. (2013). Declarative to procedural tutors: A family of cognitive architecture-based tutors. In B. Kennedy, D. Reitter & R. Amant (Eds.), *Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling and Simulation* (pp. 108-113). Centerville, OH: BRIMS Society.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249-255.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217.

Sense, F., Collins, M., Krusmark, M., Sanerson, L., Onia, J., Fiechter, J., & Jastrzembski, T. (2021). Combining cognitive and machine learning models to mine CPR training histories for personalized predictions. In *Proceedings of the 14th Educational Data Mining* (pp. 415-421). Paris, France.

Sense, F., Wood, R., Collins, M. G., Fiechter, J., Wood, A., Krusmark, M., Jastrzembski, T., & Myers, C. W. (2021). Cognition-enhanced machine learning for better predictions with limited data. *Topics in Cognitive Science*, 1-17.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*, 513-539.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive Science, 42*, 644-691.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., Pitt, M. A., & Zhou, R. (2018). Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General, 147*(9), 1325.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 171-180). Memphis, Tennessee: Springer.

Chapter 5 – Authoring Tools in Intelligent Tutoring Systems SWOT Analysis

James E. McCarthy[1] and Anne M. Sinatra[2]
[1]Sonalysts, Inc.; [2]US Army DEVCOM Soldier Center

Introduction

Over the past 50 years, the United States Department of Defense (DoD) has been a leading developer of intelligent tutoring systems (ITSs, Fletcher, 1988; Fletcher 2014; McCarthy, 2008). However, despite this long history and many demonstrable benefits, very few ITSs are currently in use within the DoD or the broader community. There are several reasons for this, but the primary one is probably the lack of authoring tools and the resultant level of effort and cost associated with their development and sustainment.

Therefore, during the 2021 Generalized Intelligent Framework for Tutoring (GIFT) Expert Workshop, leaders from the field convened to conduct a Strength, Weaknesses, Opportunities, and Threats (SWOT) analysis of ITS authoring (among other topics). This chapter summarizes those discussions, and provides recommendations for improving the GIFT (Generalized Intelligent Framework for Tutoring) software (Sottilare et al., 2017).

SWOT Analysis

During the 2021 GIFT Expert Workshop, James E. McCarthy and Neil Heffernan offered presentations that explored their experiences developing ITS authoring tools. During the presentations and the resultant discussions, participants were encouraged to offer their insights via the online *Padlet* tool (https://padlet.com/dashboard). These presentations covered the perspectives of industry and academia in addition to applications to government. This chapter summarizes those presentations and discussions. In keeping with the goals of the workshop, the presentations, *Padlets*, and chapter were organized as a SWOT analysis.

Strengths

One of the primary points of discussion was a byplay between the notion that effective authoring tools do exist and the need to define more fully what is meant by "authoring." As a community, perhaps we should more reliably distinguish among the levels and types of authoring tasks. For example, Heffernan described some of the authoring tools that he and his team have developed that allow Subject Matter Experts (SMEs; normally in the form of teachers) to build and maintain training systems (Heffernan, Authoring Tools Padlet 2021). This led to a discussion about the extent to which "content entry" qualified as authoring an ITS (Hu, Authoring Tools Padlet, 2021). The group also discussed the observation that the nature of the authoring task (and the associated interfaces for performing those tasks should be fit to the expertise of the user (Anonymous, Authoring Tools Padlet, 2021)). More generally, different users may have different levels of expertise, and may perform different authoring tasks. The authoring tool should facilitate and harmonize the contribution of each individual, independent of their personal sophistication/expertise.

Beyond the emergence of effective authoring examples, the attendees were also encouraged by the observation that the broader community is beginning to develop standards that will facilitate the authoring process. For example, in describing the Rapid Adaptive Coaching Environment (RACE) that he and his team developed, McCarthy noted that the development team used the W3C (World Wide Web Consortium) Task Model Standard (https://www.w3.org/TR/task-models/; McCarthy 2019; McCarthy, Authoring Tools Padlet, 2021). Similarly, standards for inter-simulation communication, such as IEEE Distributed Interactive Simulation (DIS; https://standards.ieee.org/standard/1278_2-2015.html) or the IEEE High Level Architecture Standard (HLA; https://standards.ieee.org/standard/1516-2010.html), enhance the visibility of the performance context and learner actions within simulation-based ITSs. Achieving this visibility is an important engineering task for these ITSs and a necessary component for associated authoring tools.

Weaknesses

After considering the strengths that enable the development of ITS authoring tools, the members of the Expert Workshop turned their attention to the weaknesses that are slowing progress.

One of the first weaknesses that the group discussed was the complement of one of the strengths. Specifically, although there *are* good examples of useful authoring systems, they are few are far between. This dearth of examples in the development system limits the ability of ideas to cross-pollinate and evolve. Graesser noted that academic work in this area is hindered by the lack of publication outlets that in turn limits the availability of empirical studies that systematically analyze the challenges and successes of human and automated authoring needed by tool designers (Graesser, Authoring Tools Padlet, 2021). Without publication venues, the academic community is discouraged from conducting the necessary foundational research.

The second weakness also reflected a topic discussed as a strength. In some of the most successful examples of authoring, the authoring task was limited to content entry. The group expressed the concern that in many other contexts, authors will need support for a more extensive list of authoring tasks. Further, the group noted that few people possess the knowledge and skills necessary to complete the range of tasks required for ITS development (Graesser, Authoring Tools Padlet, 2021). There are a couple ways to address this challenge. One approach might be labeled "Team-based Authoring." In this approach, specific tools are developed that allow experts with complementary knowledge and skills to develop their particular segment of the ITS. The authoring tool would coordinate these individual activities and consolidate them into a functional system. McCarthy (2020) labeled an alternative approach "zero authoring." In this "zero authoring" approach, flexibility is sacrificed for simplicity, reducing the required level of expertise. By pre-packing and/or parametrizing aspects of system development, authoring becomes a process of selecting specific approaches and providing the necessary parameters. As a result, while the expressiveness of the authoring tool is limited, its simplicity is increased and the authoring process is democratized.

The third primary weakness discussed by the group was closely related to the second. Specifically, despite our best efforts, the level of effort demanded even with the best authoring tools is significant. Concepts like the previously described team-based authoring and "zero authoring" would certainly address this concern. However, the group also discussed concepts such as crowdsourcing the authoring task and/or developing self-improving tutoring systems.

There was significant discussion of using platforms like StackOverflow to crowdsource the content generation and authoring portions of the development task. Tools like these provide an easy-to-use interface that allow distributed participants to answer questions and/or contribute content for use in instructional systems. The credibility of the content producers is continuously recalculated based on factors such as frequency of contribution and the frequency with which the contributor's content is used and/or endorsed (Heffernan, Authoring Tools Padlet, 2021). There was particular interest in this approach for content development. For example, Hu noted that this approach might work well for content accumulation, but expressed skepticism of its usefulness for system development and integration (Hu, Authoring Tools Padlet, 2021). Beyond that, Graesser emphasized the need for expert-provided quality control of crowd-sourced content (Graesser, Authoring Tools Padlet, 2021).

The fourth weakness that we discussed was the relatively static nature of ITSs themselves. Although these systems generally do a good job of monitoring learners and adapting to their needs, there are relatively few examples of an ITS that is designed to monitor its own actions and their success or failure with respect to promoting mastery development in learners. While conceivably this type of self-improving system could simplify the authoring task, presumably by providing guidance regarding a "good enough" initial state, the workshop developed a stand-alone chapter to this discussion (see Chapter 9; Chi et al., 2022) and we will not belabor those finding here.

Opportunities

The opportunities that were addressed within the Workshop returned us to the introduction of this chapter – the need for widespread use of a range of adaptive instructional systems, including ITSs, and the need for authoring tools that promote that level of use.

In discussing ITSs with potential users, there is very little need to convince them of the tools' usefulness. The effectiveness of ITSs in general has been studied with many different kinds of users/students (e.g., Kulik & Fletcher, 2016). Users of ITSs generally believe that the systems will work and that performance will be improved. However, they are concerned with three aspects of the ITS lifecycle. Historically, ITSs have been expensive and time-consuming to develop. Many customers could probably accept that if it was a one-time cost. However, that is often not the case. Instead, many ITSs are developed in contexts in which the target system (or its associated performance environment) undergoes frequent changes. This implies that there is a frequent (and expensive) need to update the system to keep it "current." For a smaller number of customers these costs might be acceptable. However, what is of significant concern for all users is the slowness of this update process. The timeline of the authoring process almost guarantees that an ITS will fall behind the target environment and run the risk of becoming obsolete, thus wasting the investment that the user has made. If we could reduce the cost and timeline associated with system development and maintenance, the use of these training tools would significantly increase, improving instructional efficiency and operational effectiveness.

The expense of system development stems from two sources. First, the development of these systems have traditionally required the expertise of relatively advanced cognitive scientists, software engineers, and others. The scarcity of individuals with the proper qualifications make them relatively expensive to employ. Second, the development task is largely manual and time-consuming. Together, these factors (high hourly rate and the significant level of effort) combine to make ITSs costly to develop. The former also makes development a time-consuming process. The same factors make it difficult and costly to keep deployed systems "current" as systems and procedures change. Authoring tools have the potential to address both of these factors. By "outsourcing" some level of expertise to the authoring system, these tools can reduce the levels of knowledge, skill, and experience needed to produce effective systems, opening the development task to a broader collection of individuals and reducing the "hourly cost" of the peopled involved. Similarly, authoring tools may simplify or even remove some system development steps. Doing so would reduce the level of effort associated with each step and the associated timelines. Reducing either of these factors by itself could significantly reduce the cost of ITSs. Reducing them both would have a tremendous effect. For this reason, the DoD and other users of ITSs, are tremendously interested in their development.

Although not the focus of many authoring studies or development efforts, it is important to note that authoring tools have the ability to impose a useful level of quality assurance to the development effort. Like almost any product, the quality of an ITS often reflects the knowledge and skill of its development team. While we have come to expect a significant benefit from the transition to an ITS (*e.g.,* instructional effectiveness gains of approximately one standard deviation), that gain is not a universal truth and probably reflects that most such systems result of the combined efforts of very senior/capable members of the required disciplines. Lesser teams are likely to produce inferior results. However, by "outsourcing" some of that expertise, authoring tools provide guardrails on the development process, greatly reducing the opportunity for developers to make poor decisions that can reduce system effectiveness.

Threats

Any time that a technology is emerging, care must be taken to avoid "over promising" or "over hyping" the ability of that technology. The research community must avoid making promises that do not have empirical and replicated support. The user community has limited patience for unmet promises.

Moreover, some workshop participants expressed concern that the basic authoring schemas that have been shaping our research and development efforts may be flawed (Heffernan, Authoring Tools Padlet, 2021). This threat is closely associated with concerns that we do not share definitions of learning or approaches to evaluation (Anonymous, Authoring Tools Padlet, 2021). If differences exist on such a primary level, the ability of authoring tools to produce instructional systems that promote empirically-validated learning is highly questionable. It may be wise to establish a taxonomy of definitions that the community can reference to provide an adequate context for understanding particular attempts at authoring tool development.

Overall SWOT Analysis

Table 1 illustrates the primary results of the Authoring Tool SWOT analysis.

Table 1: Overview of Authoring Tools SWOT Analysis

Strengths	1. The presence of effective exemplar authoring/maintenance tools. 2. The existence of emerging standards to facilitate authoring and inter-operation of authoring tools, engines, and tutoring systems.
Weaknesses	1. The scarcity of exemplars. 2. The scarcity of publication venues. 3. The scarcity of empirical examinations of authoring tools and processes. 4. The persistent requirement for significant expertise for tutoring system development. 5. The significant level of effort associated with most authoring tools.
Opportunities	1. The level of interest in authoring tools to reduce the cost and timeline associated with system development and maintenance. 2. The general belief that the user community has in the effectiveness of ITSs. 3. The potential for authoring tools to provide quality assurance and to guard against common authoring errors.
Threats	1. Over-promising and under-delivering can reduce the credibility of authoring system developers. 2. The schemas for authoring system development may be outdated or idiosyncratic.

Discussion and Recommendations for Future Research

The development of the Rapid Adaptive Coaching Environment (RACE) provides an interesting example of the opportunities and challenges associated with the development of authoring tools (McCarthy et al., 2019). First, as noted within the SWOT analysis, the space operations community was most interested in the development of RACE, and recognized the value of ITSs (Opportunity 2, Table 1). However, since space operations is a "fast-paced" environment characterized by frequent changes in technology, tactics, techniques, and procedures (T^3P), the stakeholders recognized the challenges associated with slow/costly system development and maintenance (Opportunity 1, Table 1). This recognition led them to explore the development of authoring tools that would allow Air Force/Space Force instructors to develop and/or maintain ITSs.

McCarthy and colleagues began their work by exploring the available literature. However, as reflected in the SWOT, the team quickly recognized that the literature was sparse (Weaknesses 1-3, Table 1) and largely subsumed three approaches:

- Machine Learning approaches to recognize "acceptable states" as part of black-box tutoring systems,
- Higher-level cognitive modeling languages to ease the process of developing a tutor's assessment logic, and
- Exemplar-based approaches that use demonstrations to define the assessment logic for procedural tutors.

The RACE team focused on the third option, largely because it harmonized well with the procedural expertise of the instructors. Consistent with our SWOT analysis, the development team's efforts were enhanced by the use of emerging technical standards (Strength 1, Table 1). For example, open interface standards allowed the team to develop an approach that allowed RACE to capture simulation events and operator actions in a generalized manner that was not dependent on a given simulation environment. Similarly, the team's use of the W3C XML (Extensible Markup Language) Task Model[1] allowed RACE to capture complex performance sequences in a manner that was independent of specific tutoring engines. Engine-specific middleware was able to ingest and apply this vendor-neutral model.

Empirical evaluations of RACE indicated that it was quite successful. Instructors with very little training could quickly produce high-quality intelligent tutoring environments (Opportunity 3, Table 1). However, the complexity of the operational domain led to an "explosion" of acceptable coaching paths. While the instructors viewed the ability of RACE to provide this coaching flexibility as a good thing, it also made them anxious about the level of effort that would be required to develop/maintain tutoring environments using RACE (Weakness 5, Table 1). Progressive refinement of the RACE authoring approach led to enhancements that emphasized simplicity and reuse, but the enhancements could never fully overcome the perceived complexity of the general-purpose RACE authoring process. As discussed earlier, this realization led the team to move toward a "zero authoring" construct in which developers sacrifice system power and generality to maximize ease of use (McCarthy, 2020).

[1] https://www.w3.org/TR/task-models/

A common finding throughout the SWOT analysis was the need for the community to openly discuss varying approaches to authoring and that the tasks that are implied within that term. It would be useful to develop, discuss, and refine a taxonomy of authoring tasks.

Another repeating theme was the need for a greater focus on, and evaluation of, authoring processes and the tools designed to facilitate that process. Workshop participants generally agreed that this line of research and development would be enhanced through the use of funded projects and appropriate publication outlets such as a design journal focused on the authoring process (Graesser, Authoring Tools Padlet, 2021). In particular, it was recognized that the community lacked a comprehensive and definitive guide to authoring within GIFT (Hoffman, Authoring Tools Padlet, 2021).

Recommendations for GIFT Overall

GIFT has a number of authoring tools that have been developed over the years. These tools provide the ability to create an entire GIFT lesson and have drag-and-drop functionality for bringing components onto a timeline. The GIFT authoring tools have been through a number of iterations with the goal of making them more user-friendly and understandable by a number of different user groups (*e.g.,* SMEs, instructors, instructional designers). While special attention was paid to the design of the question authoring system within GIFT, some of the other supporting authoring tools have either not yet been through redesigns, or still could be considered complex even after redesign (*e.g.,* Domain Knowledge File; DKF). One approach that GIFT has used is mirrored in the SWOT Analysis above, which is to try to lower the skill levels needed for the general tools such that they can be utilized by individuals without a background in computer science, while still including more advanced tools that do require extra knowledge and commitment to learn (such as authoring a DKF, which creates assessments that connect GIFT with an external software program).

While effort has gone into simplifying the general GIFT authoring tools, there is still the possibility that a new user may not initially understand how to use them, and may choose not to move forward with creating lessons. The SWOT analysis identified some approaches that authors could use within the system. For example, creating templates and example lessons that can be utilized as a starting point, and exemplars that can be edited/modified by an authoring beginner with the GIFT system. It may be beneficial to continue using a similar approach of improving the usability of all tools, while focusing most on tools with basic functionality that are likely to be used by individuals with varying backgrounds and skill levels. It may also be beneficial to complete a task analysis of the way a user interacts with the system, and determine if there are any recurrent authoring mistakes. The design of the authoring tools could potentially be updated in such a way that it reduces the likelihood of making authoring errors.

As GIFT has many use cases, and potential users, some of the lessons learned from the RACE team may be applicable. There may be benefit in building an additional simplified authoring tool similar to the "zero authoring" approach, which would allow SMEs to easily create a GIFT lesson. This approach could essentially be a simplified interface for new users who do not need a high amount of customizability in the ITSs that they are generating. The more complex authoring tool can be retained, but a new simplified one may serve as an approach to encourage authors that do not need advanced features to create lessons in GIFT.

There continue to be a number of opportunities to improve GIFT's authoring interfaces, and as additional functionality is included in GIFT authoring tools/interface design should be a consideration. There are also opportunities to examine the time it takes users to complete actions with the authoring tools, and to listen to their feedback/suggestions on ways to improve them. Work has examined the usability and functionality of more recent tools in GIFT such as the Game Master interface (Goldberg et al., 2022). GIFT has used a similar approach in the past to examine the usability of it's general tools (Ososky & Sottilare, 2016), but it has been a number of years since a similar examination has been done with the more up to date designs of the tools.

Conclusions

In most circles, the effectiveness of ITSs is not debated. Unfortunately, neither is there much debate regarding the high cost of these systems, the long duration of development efforts, and of the difficulty of keeping these systems from falling into obsolescence. Authoring tools have the potential to increase the speed and reduce the cost with which these systems can be developed and maintained. However, the lack of well-designed empirical studies and exemplar systems may be hindering our ability to develop such tools. If we overcome these challenges, ITSs will be more frequently employed, improving instructional efficiency and operational proficiency.

References

Anonymous, (2021). *Authoring Tools Padlet Comment*. URL: https://padlet.com/xiangenhu/lmsflj0gfqj2knb2

Chi, M., Biswas, G., & Hu, X. (2022). Self-Improving Systems in Intelligent Tutoring Systems SWOT Analysis. In Design Recommendations for Intelligent Tutoring Systems, Volume 10: SWOT Analysis. US Army Combat Capabilities Development Command Soldier Center. Orlando, FL, United States.

Fletcher, J. D. (1988). Intelligent Training Systems in the Military. In S.J. Andriole & G.W. Hopple (Eds), *Defense Applications of Artificial Intelligence: Progress and Prospects*. Lexington, KY: Lexington Books.

Fletcher, J. D. (2014). Digital Tutoring in Information Systems Technology for Veterans: Data Report. The Institute for Defense Analysis.

Goldberg, B., DeFalco, J.A., Hoffman, M. & Burmester, E. (2022). User Feedback on a Hybrid Team Tutoring Strategy. In Proceedings of the Challenges and Advances in Team Tutoring Workshop during AIED 2021. Available at: http://ceur-ws.org/Vol-3096/paper1.pdf

Graesser, A. (2021). *Authoring Tools Padlet Comment*. URL: https://padlet.com/xiangenhu/lmsflj0gfqj2knb2

Heffernan, N. (2021). *Authoring Tools Padlet Comment*. URL: https://padlet.com/xiangenhu/lmsflj0gfqj2knb2

Hoffman, M. (2021). *Authoring Tools Padlet Comment*. URL: https://padlet.com/xiangenhu/lmsflj0gfqj2knb2

Hu, X. (2021). *Authoring Tools Padlet Comment*. URL: https://padlet.com/xiangenhu/lmsflj0gfqj2knb2

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. Review of educational research, 86(1), 42-78.

McCarthy, J. E. (2008). Military Applications of Adaptive Training Technology. In M.D. Lytras, D. Gašević, P. Ordóñez de Pablos, & W. Huang (Eds.), *Technology Enhanced Learning: Best Practices*. Hershey, PA: IGI Publishing.

McCarthy, J. E., Kennedy, J., Grant, J., & Bailey, M. (2019, July). Developing Authoring Tools for Simulation-Based Intelligent Tutoring Systems: Lessons Learned. In *International Conference on Human-Computer Interaction* (pp. 118-129). Springer, Cham.

McCarthy, J. E. (2020, July). Toward Zero Authoring: Considering How to Maximize Courseware Quality and Affordability Simultaneously. In *International Conference on Human-Computer Interaction* (pp. 144-163). Springer, Cham.

McCarthy, J. E. (2021). *Authoring Tools Padlet Comment*. URL: https://padlet.com/xiangenhu/lmsflj0gfqj2knb2

Ososky, S. J., & Sottilare, R. A. (2016). *A heuristic evaluation of the generalized intelligent framework for tutoring (gift) authoring tools*. US Army Research Laboratory Aberdeen Proving Ground United States.

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

Chapter 6 – Domain Modeling in Intelligent Tutoring Systems SWOT Analysis

Vasile Rus
University of Memphis

Introduction

This chapter presents an overview of domain models and domain modeling techniques and methods with a focus on their application in the area of adaptive instructional technologies. Furthermore, implications and recommendations for future educational technologies including the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare, Brawner, Goldberg, & Holden, 2012) are being discussed.

Domain modeling is the task of specifying the units of knowledge, also called Knowledge Components (KCs), in a target domain such as physics, biology, or computer programming. More specifically, a domain model includes a structure that specifies the relationship among the KCs, typically in the form of a prerequisite knowledge structure suggesting a specific trajectory towards mastery, i.e., a particular order in which students should master the KCs (Goldin, Pavlik Jr, & Ritter 2016; Koedinger, Corbett, and Perfetti 2012; Chau et al. 2020). A domain model can also link the KCs to specific learning activities or objects that allow learners to master those KCs through practice. We can make an argument that domain modeling should be expanded to include all key concepts, skills, ideas, principles, other types of knowledge such as procedures and processes, and the values, identity, and epistemology of the community of experts or professionals active in the target domain. That is, if the goal of instruction is to prepare a successful expert in a domain, besides the KCs in textbooks, a learner must learn, for instance, the values of the experts in the target domain and therefore domain models must specify those additional aspects of becoming an expert in a community of experts.

A domain model is the outcome of the domain modeling process which can be manual, semi-automatic, or fully automated. We will address some key issues related to domain models and the authoring process of such domain models which is a key step in developing adaptive instructional systems (AISs).

Indeed, the domain model is one of the key components of AISs besides the pedagogical model, the learner model, and the interface or interaction model (Sottilare et al., 2012). From an AISs architectural perspective, the domain model should provide to the other key components all the necessary information about the domain: (i) provide the pedagogical model with links from KCs to specific learning activities that give learners the opportunity to master those KCs through practice (Are the links part of the domain model? Are the learning activities part of the domain model?); (ii) provide the learner model a list of KCs which the assessment module must evaluate and update the learner model accordingly; and (iii) provide the interface model with information for visualization and authoring/editing of the domain model, visualize the learner model as an overlay model over the domain model which is used in some cases, etc. The interdependence and interplay among the key components of an AIS can be quite complex. The role of the domain model cannot be overstated as it affects almost any aspect of AISs. As an example, the pedagogical model may select the next instructional task and the best instructional strategies for a given student for a target KC provided by the domain model. Many times the strategy may need to be adapted to the target domain or even a particular KC. This adaptation is based on pedagogical content knowledge, a well known area of research in education (PCK; Shulman, 1986), which implies that domain models should include such pedagogical content knowledge as well.

This chapter will present a brief overview of domain models and of prior work followed by an example of a domain model and of a domain model extraction technique for building adaptive instructional systems for intro-to-computer programming.

What is A Domain Model?

Given that the goal of learning is to master a target domain, the first step of any learning effort must start with specifying what is to be learned, i.e., a specification of a domain or domain model. Domain models have been defined in various ways, some more comprehensive than others and often guided by some underlying theories or frameworks such as the Knowledge-Learning-Instruction framework that proposes decomposition of knowledge into knowledge components (abbreviated as KCs) (Koedinger et al., 2012). KCs may include what others call skills, concepts, schemas, or other labels. In this chapter, we will use the term KC and define it as an atomic unit of knowledge which cannot be decomposed anymore at least from a particular domain perspective. It should be noted that diversity of contexts in which a KC may occur may lead to a (very large/potentially infinite) number of nuanced KCs. These different incarnations with subtle differences among them may or may not be regarded as new/more specific KCs (i.e., the atomic nature may not seem atomic anymore simply because of contextual differences). The small differences in different contexts may imply the use of different instructional strategies and trigger different misconceptions.

Definitions vary from focusing on key concepts to be mastered in a domain and their prerequisite structure as indicated by Pelánek (2020, pp. 535), "domain modeling - designing an appropriate organization of individual learning objects to higher-level units and specification of relations among these units." to more comprehensive definitions as the one provided by Pavlik and colleagues (2013, pp. 39): "The domain model contains the set of skills, knowledge, and strategies of the topic being tutored. It normally contains the ideal expert knowledge and may also contain the bugs, mal-rules, and misconceptions that students periodically exhibit. It is a representation of all the possible student states in the domain. While these states are typically tied to content, general psychological states (e.g., boredom, persistence) may also be included, since such states are relevant for a full understanding of possible pedagogy within the domain."

These definitions seem to focus on the key concepts of the domain ("key concepts", "the topic being tutored") and less so on other important aspects of being an expert in a field such as values and epistemology of experts in the domain. Furthermore, those definitions do not mention pedagogical content knowledge or links to pedagogical content knowledge in case such knowledge is embedded in the pedagogical model of an AIS as opposed to it being embedded in the domain model. Therefore, we propose a new, more comprehensive definition of domain models that tries to capture all that is needed to become a successful expert in a domain: a domain model should include all key concepts, skills, ideas, principles, as well as other types of knowledge such as procedures and processes and the values, identity, and epistemology of the community of experts or professionals active in the target domain. While cognitive modeling of a domain has been around for some time, the role of values, identity, and epistemology in mastering a domain has been studying more recently (Bagley & Shaffer, 2009). Furthermore, domain models should include other relevant knowledge such as pedagogical content knowledge or links to such knowledge if that knowledge is embedded somewhere else, e.g., the pedagogical model.

When used in AISs, the underlying framework shapes the way the domain model is represented. For instance, in cognitive tutors (Koedinger et al., 2004) the domain model is represented mainly as a prerequisite knowledge structure with nodes representing the KCs and the edges representing some order or prerequisite relations among the KCs. In constraint-based tutors (Mitrovic et al., 2003), the domain model is represented as a set of constraints whereas in model-based tutors (Kumar, 2002) the structure and behavior of the domain is captured. In conversational tutors (Rus et al., 2013), the knowledge is represented in the form of natural language statements called expectations (correct knowledge) and misconceptions (frequent pitfalls that beginners experience).

Once a representation has been chosen for the domain, the next step is to do an initial domain specification. This can be done by an expert or it can be inferred automatically from sources such as existing textbooks followed by a quality assurance step done by an expert. Initial domain specifications are then constantly refined based on student performance data (Fancsali & Ritter, 2020).

Domain models can vary in their complexity from a list of key concepts to be mastered, e.g., extracted from the glossary or table of contents of a textbook, to more refined domain models based on student performance data to personalized domain models such as models that are tailored to a particular student. The latter, student-specific domain models may rely on performance of a 'similar' student in the past and/or based on the current students' performance so far. Table 1 presents a brief SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis of domain models and domain modeling.

Table 1. A brief SWOT analysis of domain models and domain modeling for developing AISs.

Strengths	Weaknesses	Opportunities	Threats
● *Domain experts and expertise, e.g., encoded in textbooks that can be tapped into (directly and indirectly)* ● *Data is available for established domains,* ● *Commercial and academic/research domain models being developed* ● *Emergence of novel, data-driven, semi-automated authoring processes for domain modeling* ● *Initial domain model inference methods* ● *Domain model refinement methods* ● *Fine-grain student performance data collected at scale* ● *Standardization efforts emerging*	● *Conceptualization* ● *Lack of ready to use domain models for AIS development* ● *Imperfect automated methods (i.e., Natural Language Processing (NLP))* ● *Lack of guidance and agreements with regard to which domain modeling parts are/should be shareable and which proprietary* ○ *Skeleton + proprietary approach* ● *No widely acceptable, domain model authoring process* ● *Not enough educators/teachers' involvement/adoption*	● *Better define domain models and their role in AISs and in the learning ecosystem as well as their success metrics* ● *improve and increase automation based on data-driven methods* ● *Advances in AI/ML/Data Science/NLP promise (semi-)automation of the domain modeling authoring/refinement* ● *Standardization for well-established domain* ● *Skeleton + proprietary approach* ● *AISs/EdTech/Data Science literacy for educators* ● *Much overlap with learner modeling* ● *Solutions for secure and privacy-preserving data access and sharing*	● *Conceptualization* ● *Data access (privacy and security)* ● *Slow progress on open vs. proprietary aspects of domain models* ● *Authoring is (still) expensive*

Among the major Strengths of state-of-the-art domain modeling would be the increasing data available in electronic form (e.g., textbooks, student performance data) which can be used to infer and refine domain models. Among the weaknesses, we note a relatively weak definition of what a domain model is. This chapter is meant to address this weakness by proposing a more comprehensive definition. This weakness is also a potential threat as a poor conceptualization of what a domain model should be could have negative ripple effects on other components of an adaptive instructional system. Among opportunities, we emphasize the role of advanced in Artificial Intelligence/Machine Learning/Data Science/NLP methods that hold the promise of (semi-)automating the domain modeling authoring and refinement process. Furthermore, we suggest the adoption of a skeleton+proprietary approach as a way to make progress towards standardization. In this approach, there will be a skeleton part (shared/open part of a domain model) as well as a vendor specific part or proprietary part of a domain model. This approach should offer a good trade-off between the need for standardization and the need of vendors to keep their 'secret sauce' secret. Among threats, we highlight data privacy, security, and ownership. While data is available, access to data and in particular fine-grained, student performance data is still a challenge. Solutions that offer a compromise between the need to infer patterns and trends in the data while preserving privacy and ownership are needed.

Prior Work

We briefly review prior efforts related to automated extraction of domain models and the related area of automatic domain model refinement.

When developing domain models, there are three significant information sources: experts, textbooks (written by domain experts), and learner data. We highlight work focusing primarily on extracting or refining domain models from data (text or structured data). For instance, student performance data is often used as input to domain modeling methods in the form of a Q-matrix linking KCs to instructional items in a domain, such as solutions to problems, steps in a solution, or a student explanation.

Such Q-matrices are useful primarily for well-defined domains and less so for ill-defined domains (Goldin, Pavlik Jr, & Ritter 2016). Given such a Q-matrix, one can infer a set of latent variables that can partition a set of instructional items based on learner responses to those items. Prediction of student performance based on the discovered latent skills is used to evaluate the inferred domain model. There are several issues with such approaches to domain model discovery: (1) interpreting what skills the latent variables represent and (2) the need for student performance data. The latter is quite challenging when developing domain models for emerging domains such as data science or nanotechnology for which student data may not yet be available. Often Q-matrix-based approaches start with a domain model (original model) which is another challenge as they require some other source of the original or start domain model. The main goal in such cases is to refine the start domain model based on student performance data, i.e., discovering a new set of skills in the form of latent variables that best predict student performance.

Extracting a start domain model (initial version of a domain model) automatically has been explored before through information extraction (IE), a major subarea of Natural Language Processing (NLP), from textual sources such as online data sources that can also greatly help in understanding and refining existing KC structures. For example, consider a potential statement from Wikipedia such as "using factorization we can solve quadratic equations." (not actual quote) From the text in the statement, one can easily infer that factorization is related to understanding how to solve quadratic equations. This can be especially useful for discovering new concepts as well as refining existing concepts encoded in systems such as MATHia (Fancsali & Ritter, 2020).

Extracting KCs from textual sources such as textbooks has been explored. For instance, Chau and colleagues (Chau et al., 2020) adopted a supervised machine learning (ML) approach based on a set of expert-defined features. The features they used fall into three broad categories: linguistic, positional, or statistical. They used an over-generation and ranking approach. They first generated a large set of candidate keyphrases and then applied a selection criterion or filter to rank and detect the true domain concepts. They compared their approach to a number of baseline methods and some off-the-shelf algorithms such as TextRank (Mihalcea & Tarau, 2004).

A significant effort has been put in scaling and improving data-driven domain models for cognitive tutors, i.e., cognitive model refinement. The goal of ongoing efforts such as the NSF-funded Learner Data Institute project (LDI; Rus et al, 2020; Fancsali & Ritter, 2020) is to scale, improve, and extend data-driven methods to validate and refine KC models and also seeks to develop automated methods that will help to alleviate the cost of developing such KC models in the first place. Learning Factors Analysis (LFA; Cen et al., 2006), a semi-automated approach to refine cognitive KC models, has been shown to improve the statistical accuracy of models learned across datasets and domains (Koedinger et al., 2012). LFA is an approach that iteratively searches over alternative KC models by considering situations in which KCs are "split" into two or more KCs and two KCs are "merged" to become a single KC. For example, KCs can be "split" into new KCs according to difficulty factors that may pertain to a student's ability to solve an element of a problem. For instance, it may be the case that students find it more difficult to find the area of a circle that is inscribed in a square. A hypothesized cognitive model may only contain the KC "find the area of a circle," but empirically we can test whether such a KC is appropriate or whether we should "split" this single KC into two, one corresponding to finding the area of a circle when it is pictured alone and another corresponding to when it is embedded within another shape (Cen et al., 2006). Discovered models not only improve in terms of statistical accuracy but have also been found in a small-scale "close-the-loop" study to lead to more effective tutoring, resulting in improved learning efficiency and superior student learning gains. LFA allows the researcher to generate new hypotheses about the KCs underlying a domain and statistically test them to validate whether they constitute genuine improvements over previous models. The LDI team is working, for instance, on optimizing and scaling up LFA. LFA search (Koedinger et al., 2012) has led to the discovery of substantially improved models, relative to Subject Matter Expert (SME)-coded KC models. These results, however, are limited to a set of eleven datasets, each of which is relatively small, ranging in terms of number KCs from 1 to 48 and in the number of student users from 41 to 318. The LDI team is currently working (Fancsali & Ritter, 2020) towards scaling up the LFA search for improved cognitive models to a much larger corpus of data in terms of the number of students, student problem solving actions, and broader coverage of full curricula (i.e., increased numbers of KCs as well as doing many searches to refine the KC model for each topic unit in the curricula, which are modeled and tracked independently as students work through them). As in Koedinger et al. (2012), we expect this LFA search to not only result in improved KC models but to suggest changes to the content and structure of the problems given to students. One benefit of interpretable KC models is that they afford such analysis.

Furthermore, the LDI team of which the author is part of is working on explainable KC models, i.e., on developing advanced methods to improve KC structure by seeking input from more powerful modeling frameworks that enable explainable statistical models. As already mentioned, one of the core tasks in domain modeling is to determine when to split/combine KCs. ML methods can find statistical patterns in data that can help determine splitting/combining criteria, but several of these methods (e.g. deep learning) are inherently non-interpretable/explainable. In the absence of explanations, performing splits or combinations simply based on the output of a machine learning algorithm may seldom lead to true learning gains. Therefore, some LDI team members including the author of this chapter are working on an explainable framework based on a neuro-symbolic approach that combines SRL (statistical relational learning), Markov Logic Networks (MLNs), and probabilistic soft logic (PSL) with deep neural networks. Our proposed approach is to learn the explanation for a KC as a (bounded) subset of MLN/PSL formulas. To do this, we aim to model the dependencies between KCs and problem difficulty as MLN/PSL formulas. Based on the data (student performance), we perform probabilistic inference and obtain a set of weighted formulas that best explain inference results on a specific KC. A domain expert can then verify this explanation for a KC and decide on the split/combine decisions of KCs. A key technical challenge in this task is to tractably compute the explanations across hundreds or thousands of KCs, when the MLN can contain hundreds of thousands of possible instantiations (depending on the amount of student data). To do this, we utilize advanced approximate counting based sampling and local-search methods (Venugopal & Rus, 2016) and exploit parallel computing frameworks such as Spark (Cheekati et al., 2016) to obtain scalable explanations.

An Example

In this section, we highlight a domain model for intro to programming and a domain modeling method that automatically discovers the domain model by extracting key concepts from intro-to-programming textbooks (Banjade et al., 2021). The work was done in the context of developing and investigating the effectiveness of an Intelligent Tutoring System (ITS) for source code comprehension.

As already noted, a typical ITS works with students through various instructional activities to help them master key concepts of a target domain. The underlying domain model guides the functionality of ITSs and has a major impact on the system's effectiveness to induce learning gains and on the overall student learning experience.

The key concepts in a target domain that students need to master are often specified by experts such as domain experts, pedagogical experts, and ITS designers. This expert-driven approach is tedious, expensive, time-consuming, and makes ITS development hard to scale across domains. Furthermore, expert-defined domain models can be error prone or inadequate for instructional purposes as "experts may forget the difficulties that novice learners face." (Goldin, Pavlik Jr, & Ritter 2016; pp: 115). This can have negative consequences on assessing learners' knowledge state, which leads to poor adaptivity of ITSs and, consequently, a negative impact on the effectiveness and overall quality of the provided instruction.

To overcome the above-mentioned challenges, there is a need for automated or semi-automated methods. We highlight here a novel automated method for domain model discovery, particularly focusing on computer programming textbooks. Such an automation has several advantages. First, it relieves the need for handpicking key concepts as the textbook authors already put much effort in doing so. Second, it helps discover the ordering of key concepts necessary for tutoring systems as textbooks present the key concepts in a particular order (which could be refined based on, for instance, student performance data). Third, automating knowledge discovery from the textbooks will save a lot of time and effort for tutoring system developers. In particular, it will help with porting an ITS platform from one domain to another more easily thus leading to more scalable ITSs across topics and domains.

Our approach to automatically extract domain models from textbooks was to rely on keyphrase extraction methods to identify a domain's key concepts. The problem of extracting key concepts from a Computer Science textbook poses several unique challenges and opportunities. For instance, Computer Science textbooks contain domain-specific words such as *for* to describe the concept of loops, and therefore this key phrase requires special handling to distinguish it from the regular preposition "for". Furthermore, Computer Science textbooks contain many code examples and their plain text explanations, so, there is a practical need for distinction between the two. Typical key phrase extraction methods work primarily on pure text. This combination of code and text in Computer Science textbooks is also a great opportunity for domain modeling as it facilitates the linking of key concepts to specific learning activities such as code comprehension activities.

For instance, a Java code example in an intro to Java programming textbook can be linked to the key concepts it covers by inspecting the key concepts mentioned in the explanatory text. Furthermore, intro to programming textbooks document major misconceptions students exhibit while learning programming. Our goal is to expand a typical domain model with the key misconceptions students have, which is critical for feedback opportunities in ITSs. In sum, the proposed method for automated discovery of domain models addressed the following four key tasks: (1) knowledge component extraction, (2) prerequisite knowledge structure discovery (3) linking of key concepts to learning objects/activities, which in our case, a knowledge object is a Java example in the textbook, and (4) misconception extraction.

We adopted an over-generation and ranking approach to discover the key concepts of the intro to programming domain. Our inputs are Computer Science textbooks. There are several advantages of using textbooks to extract domain models. First, textbooks describe a target domain's knowledge with an instructional purpose in mind. The authors of textbooks spend significant efforts to define the key concepts, present them in a specific order, and provide plenty of instructional activities to practice those concepts. Furthermore, the textbooks' structure in chapters and sections facilitates the extraction of key concepts using, for instance, statistical methods. It enables the organization of the extracted key concepts in more complex structures such as prerequisite knowledge structures and taxonomies.

Furthermore, as already noted, intro to programming textbooks have a peculiarity in that they contain both code examples and related explanatory text. Since the main objective was to extract the key concepts, the focus was only on the text explanations instead of code examples. The code examples generally contain comments in text form that explain the code as well. Often, those comments repeat concepts described in the surrounding explanatory text and are therefore redundant for our purposes. It is possible to extract more abstract concepts directly from code, e.g., by performing a static syntax analysis of the code but is beyond the scope of this chapter. It should be noted that there is a major disadvantage of such methods - the extracted concepts are harder to interpret. For these reasons, we focused primarily on extracting the text portions of intro to programming textbooks. To extract the descriptive text from textbooks, we developed a Naive Bayes classifier that can classify each line in textbooks as either explanatory text or code. This classifier had a classification accuracy of 94% with *F*-score of 0.9. The explanatory text, thus extracted is used for further analysis. We used Introduction to JAVA programming (Liang, 2011) for our experiments. We evaluated statistical and graph-based methods for domain model extraction for the target domain of intro to computer programming and obtained recall as high as 0.60 and precision as high as 0.75. We measured precision and recall at various ranks 1, 10, 20, …, and 100, i.e., precision and recall at rank 1, precision and recall for top 10 ranked candidate key concepts, top 20, and so on. The results suggest that unsupervised key phrase extraction methods can be used for domain model discovery from Computer Science textbooks.

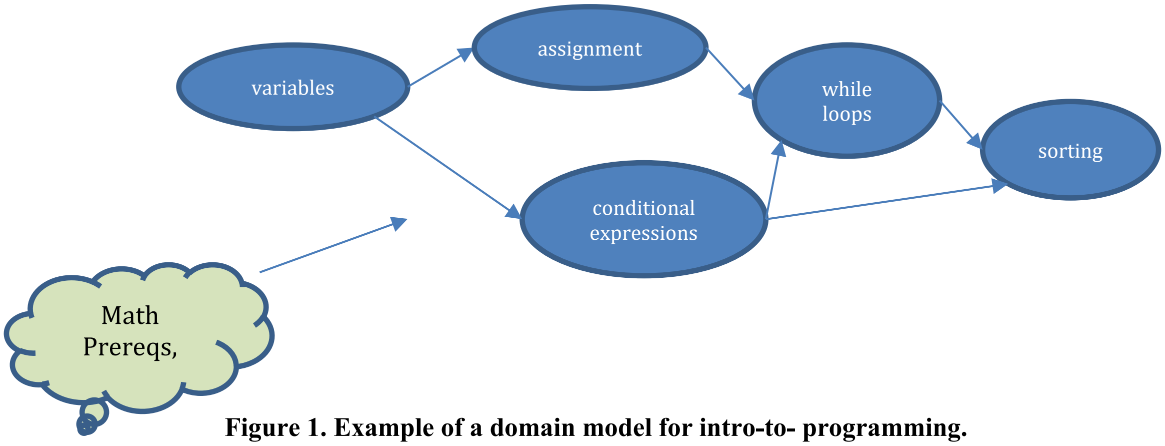


Figure 1. Example of a domain model for intro-to- programming.

Recommendations and Future Research

Using VanLehn's two-loop framework for AISs (VanLehn, 2006), the domain model drives primarily the outer loop, i.e., the domain model's main role is to suggest potential trajectories towards mastery over the key concepts or knowledge components in a domain, i.e., a particular order in which students should attempt to master the KCs. This implies and also explains the prerequisite structure over KCs that domain models typically include. Nevertheless, the domain models are also needed for *the inner loop* because the domain models include misconceptions which are detected (and corrected immediately) during the inner for fully-adaptive AISs. It should be noted that there are AISs that are not fully adaptive, e.g., they may include only macro-adaptivity, i.e., only outer-loop adaptivity in terms of selecting the best sequence of instructional tasks and the corresponding KCs. Such macro-adaptive only systems have no inner-loop and therefore no micro-level or within task adaptivity. That is, while the student is working on a given instructional task their performance is not monitored and no feedback is provided at step level. Feedback is only offered at macro-level, i.e., whether the student successfully finished the task or not. An examples of such a system is ALEKS (Falmagne & Doignon, 2011). Such macro-adaptive systems therefore need less sophisticated domain models compared to fully-adaptive AISs. This is important to keep in mind for developers in general and for GIFT, that is, depending on the goal and characteristics of the AISs, a more or less sophisticated process and end product for domain modeling is needed. The implications and recommendations for GIFT are therefore to include features that enables developers of AISs to implement and specify domains models that may serve the outer loop, the inner loop, or both. Other purposes should be kept in mind such as assessment and reporting of individual and group student performance, for instance.

Conclusions

This chapter provided a brief overview of domain models and of prior work followed by an example of a domain model and of a domain model extraction technique for building AISs for intro to computer programming. We provided what we believe is a more comprehensive definition of a domain model to include all that is necessary to prepare a successful expert in a domain, i.e., besides all key concepts a learner must learn the skills, ideas, principles, other types of knowledge such as procedures and processes and the values, identity, and epistemology of the community of experts or professionals active in the target domain. This more comprehensive definition should lead to more comprehensive domain models which in turn should lead to more effective tutors.

Acknowledgements

This research was partially sponsored by the National Science Foundation under award The Learner Data Institute (award #1934745) and an award from the Department of Defense (U.S. Army Combat Capabilities Development Command - Soldier Center). The opinions, findings, and results are solely the authors' and do not reflect those of the funding agencies.

References

Bagley, E., & Shaffer, D. W. (2009). When people get in the way: Promoting civic thinking through epistemic gameplay. *International Journal of Gaming and Computer-mediated Simulations*, *1*, 36-52.

Banjade, R., Oli, P., Tamang, L. J., Chapagain, J., & Rus, V. (2021). Domain Model Discovery from Textbooks for Computer Programming Intelligent Tutors. The International FLAIRS Conference Proceedings, 34. https://doi.org/10.32473/flairs.v34i1.128561

Cen, H., Koedinger, K., Junker, B. (2006). *Learning factors analysis – a general method for cognitive model evaluation and* *improvement*. In Proceedings of the Eighth International Conference on Intelligent Tutoring Systems (ITS2006) (pp. 164-175) Berlin: Springer.

Chau, H., Labutov, I., Thaker, K., He, D., and Brusilovsky, P. (2020). Automatic concept extraction for domain and student modeling in adaptive textbooks. International Journal of Artificial Intelligence in Education 1–27.

Cheekati, H.C., Goli, S., & Venugopal, D. (2016). MSpark: A Scalable Lifted Inference Pipeline for MLNs, In IJCAI-16 workshop on Statistical Relational Artificial Intelligence, 2016.

Fancsali, S.E. & Ritter, S. (2020). Data-Intensive Learning Engineering & Applied Education Research with Carnegie Learning's MATHia Platform, In V. Rus & S.E. Fancsali (Eds.) Proceedings of The First Workshop of the Learner Data Institute - Big Data, Research Challenges, & Science Convergence in Educational Data Science, The 13th International Conference on Educational Data Mining (EDM 2020), July 10-13, Ifrane, Morroco (held online).

Falmagne, J.C. & Doignon, J.P. (2011). Learning Spaces. Springer-Verlag, 2011.

Koedinger, K.R., Aleven, V., Heffernan, N., McLaren, B., and Hockenberry, M. (2004). Opening the Door to Non-programmers: Authoring Intelligent Tutor Behavior by Demonstration. in 7th Int. Conf. Intelligent Tutoring Systems. 2004. Maceio, Brazil: Springer-Verlag. p. 162-174.

Koedinger, K. R., Corbett, A.T., and Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive science36(5):757–798.

Koedinger, K.R., McLaughlin, E.A., Stamper, J.C. (2012). Automated student model improvement. In Proceedings of the Fifth International Conference on Educational Data Mining (EDM2012) (pp. 17-24) International Educational Data Mining Society.

Kumar, A.N. (2002). Model-Based Reasoning for Domain Modeling in a Web-Based Intelligent Tutoring System to Help Students Learn to Debug C++ Programs. Intelligent Tutoring Systems.

Liang, D. (2001). Introduction to java programming. Prentic Hall, 2001.

Goldin, I.,Pavlik Jr, P. I., and Ritter, S. (2016). Discovering domain models in learning curve data. Design Recommendations for Intelligent Tutoring Systems, 115.

Mihalcea, R., and Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 404–411.

Mitrovic, A., Koedinger, K. R. & Martin, B. (2003). A Comparative Analysis of Cognitive Tutoring and Constraint-Based Modeling. In *User Modeling 2003* (Vol. 2702/2003): Springer Berlin / Heidelberg.

Pavlik, P.I., Brawner, K.W., Olney, A.M., & Mitrovic, A. (2013). A Review of Student Models Used in Intelligent Tutoring Systems, In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), Design recommendations for adaptive intelligent tutoring systems (Learner modeling, Vol. I, pp. 39–68). Orlando: US Army Research Laboratory.

Pelánek, R. Managing items and knowledge components: domain modeling in practice. Education Tech Research Dev 68, 529–550 (2020). https://doi.org/10.1007/s11423-019-09716-w

Rus, V., D'Mello, S., Hu, X., & Graesser, A.C. (2013). Recent Advances in Conversational Intelligent Tutoring Systems, AI Magazine, 34(3):42-54.

Rus, V., Fancsali, S.E., Bowman, D., Pavlik Jr., P., Ritter, S., Venugopal, D., Morrison, D., & The LDI Team. (2020). The Learner Data Institute: Mission, Framework, & Activities. In V. Rus & S.E. Fancsali (Eds.) *Proceedings of The First Workshop of the Learner Data Institute - Big Data, Research Challenges, & Science Convergence in Educational Data Science*, The 13th International Conference on Educational Data Mining (EDM 2020), July 10-13, Ifrane, Morroco (held online).

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.

Sottilare, R.A., Brawner, K.W., Goldberg, B.S., & Holden, H.K. (2012). *The Generalized Intelligent Framework for Tutoring* (GIFT). Downloaded from www.gifttutoring.org on November 30, 2012.

Venugopal, D. & Rus, V. (2016). Joint Inference for Mode Identification in Tutorial Dialogues, Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), Osaka, Japan, December 2016.

VanLehn, K. (2006). The behavior of tutoring systems. International Journal of Artificial Intelligence in Education. 16(3), 227-265.

Section III – Advanced Elements of Intelligent Tutoring Systems SWOT Analyses

SWOT Analyses of:

Assessment

Team Tutoring

Self-Improving Systems

Data Visualization

Competency-Based Scenario Design

Chapter 7 – Assessment in Intelligent Tutoring Systems SWOT Analysis

Diego Zapata-Rivera[1] and Xiangen Hu[2]
Educational Testing Service[1;], University of Memphis[2]

Background

As educational assessment systems become more interactive and technology-rich, core values of educational assessment such as validity, reliability, comparability, generalizability, and fairness are expected to continue playing an important role in the design and evaluation of educational assessments (Mislevy, 2021). These properties also apply to those assessments embedded in Intelligent Tutoring Systems (ITSs).

Different types of assessments have been designed for different purposes (e.g., summative assessments, formative assessments, and integrated assessment systems). Understanding the characteristics of each type of assessment and the context in which it will be used can facilitate the process of selecting the right type of assessment for the right purpose. Misalignment between assessment type and expected uses usually generate confusion and frustration. Some common misalignment issues include selecting a summative assessment to provide immediate instructional feedback (e.g., next steps to inform teaching and learning) or using a census model, where every member of the population takes the assessment, when a sampling approach would suffice to provide aggregate level results for policy makers (Mislevy, 2019).

The current state of assessment is characterized by comprehensive assessment systems that make use of summative and formative assessments. These assessment systems may include the use of different types of tasks (e.g., multiple choice, constructed response, and simulations) in performance-based assessments that are used to assess content areas, domain practices, and 21st century skills (e.g., collaborative problem solving). They can take the form of integrated assessments that assess several skills using a single task or scenario (e.g., assessing both listening comprehension and speaking or writing in a technology rich English language task). Some of these assessments make use of process and response data and employ automated scoring (e.g., AI scoring engines). These assessment systems provide support for accessibility and accommodations.

Advances in educational assessment, cognitive science, and artificial intelligence have made it possible to integrate valid assessment and instruction in the form of modern ITSs (Shute & Zapata-Rivera, 2010, 2012). ITSs gather evidence of learners' knowledge, skills, and other attributes (KSAs) to provide various types of support (e.g., adaptive feedback and adaptive sequencing of tasks). ITSs may include game elements, dialogue systems, and interaction with virtual agents. ITSs can gather a variety of learner performance data including both process and response data.

In this chapter, we elaborate on emerging trends in educational assessment. Also, we illustrate some of these trends in the context of "caring" assessment. Next, we present an analysis of the strengths, weaknesses, opportunities, and threats (SWOT) of assessment in ITSs. This analysis builds on reports describing assessment issues in ITSs as well as personal communications with researchers in the areas of assessment and ITSs. Finally, we elaborate on future work and provide recommendations for improving assessment in the Generalized Intelligent Framework for Tutoring (GIFT, Sottilare, et al., 2017).

Considerations about the Future of Assessment

Bennett (2018) indicates that future assessments will be technology-based, measure new constructs, be built from richer underlying models of cognition and learning, make greater use of more complex and personalized tasks that attempt to improve learning, be better at accounting for context, be embedded and distributed across time, use automated scoring, incorporate new approaches to modeling and analysis, and provide more effective reporting.

Several of these assessment trends are also mentioned by Mislevy (2019, 2021) who also emphasizes assessments that are more closely related to learning contexts. He describes assessment as an argument-structured, contextualized process of evidentiary reasoning, with measurement machinery as part of the toolkit supporting its application. Thus, measurement remains important as a framework for evaluating the quality and quantity of information and as metrics for improving learning and assessment processes. Other aspects of future work include exploring the interplay of Bayesian inference and data-analytic methodologies, integration of general frameworks (e.g., from psychological research and domain-based research), and considering tradeoffs between local usefulness and broader comparability of assessments.

Shute (2016, 2019) envisions a continuous assessment approach where technology rich environments include embedded innovative tasks designed using principled assessment design (e.g., Evidence-Centered Design, or ECD; Mislevy, Almond, & Lukas, 2003). These embedded tasks will be pedagogically relevant and provide socially and emotionally meaningful learning situations for students. These environments will provide useful feedback during the learning process (Shute, Hansen, & Almond, 2008). Lastly, she expects the future will bring more work on assessments that blur the line between what can be considered to be formative or summative.

There are other similar visions that integrate learning processes and assessments in in digital environments. Baker (2019) forecasts the use of more performance-based assessments that measure multiple constructs at once and are used in the context of learning. The use of machine-learned computational models for specific tasks, a focus on "ill defined" constructs (e.g., emotion, collaboration), more applications of advances in speech and language processing for discourse analysis, and the use of multiple, multimodal sources of process and response data (e.g., via sensors) (D'Mello, Gregg, & Southwell, 2020; D'Mello, Tay, & Southwell, 2022).

In general, assessment trends include:

- Increased attention to and expansion of *contexts* in assessments.
- Assessment that supports the learning process.
- Tradeoffs to balance local usefulness and broader comparability.
- Performance-based, technology rich environments that can measure multiple constructs at once.
- Innovative assessments developed using principled assessment design.
- Models that make use of process and response data.
- Robust multimodal sensing in context.

The Case of Caring Assessment

The work on caring assessments (CAs) illustrates some elements of future assessments (Zapata-Rivera, 2017; Zapata-Rivera, Lehman, & Sparks, 2020). Caring assessments consider additional information about the learner and learning context to create situations that learners find engaging and at the same time can be used to collect valid and reliable evidence of learners' KSAs. The learner model in caring assessment includes cognitive, social, emotional, and other characteristics of the learner and the learning context that may influence levels of engagement and performance (e.g., socio-cultural or motivational aspects).

By expanding the scope of the learner model, it is possible to design a variety of adaptations aimed at supporting learners. These adaptations may include accessibility support, adaptive feedback and sequencing of activities, gathering additional evidence of learners' KSAs using conversations and other activities, granting additional time, giving opportunities to make revisions, recommending materials/activities, and making changes to administration conditions.

Some recent work in this area includes (a) exploring the role of emotions in CAs (Lehman & Zapata-Rivera, 2018) and (b) exploring individual differences in CAs (Sparks et al, 2018; 2019). work on the role of emotions in CAs involves detecting emotions, tracking and responding to emotions, and examining the impact of emotions (e.g., emotion type and intensity) on the quality of the evidence gathered in conversation-based tasks (Lehman & Zapata-Rivera, 2018; Lehman, Sparks & Zapata-Rivera, 2018). Work on individual assessments in CAs explores the quality of responses based on variables such as grade level, opportunity to learn, personality variables, socio-emotional skills, and other "non-cognitive" characteristics related to achievement (e.g., self-efficacy; persistence, growth mindset, cognitive flexibility, and test anxiety). These lines of work can result in insights for the type of support needed for particular subgroups of learners in order to maintain and improve engagement and support learning.

We are also looking at how to apply cognitive bandwidth recovery strategies to improve assessment of underserved students (Verschelden, 2017). Cognitive bandwidth recovery strategies are designed to minimize the negative impacts of cognitive resources dedicated to dealing with the effects of poverty, racism, and social marginalization. These strategies promote a growth mindset and self-efficacy. Some of these strategies have potential for improving the chances of minority students to demonstrate what they know or can do, which in turn can improve their educational opportunities through more accurate assessment of their knowledge and skills.

SWOT Analysis

In this section we describe the results of a SWOT analysis of assessment in ITSs. Information from this analysis was taken from research reports describing different aspects of assessment in ITSs (e.g., Conati, 2009; Katz et al., 2017; Mislevy & Yan, 2017; Shute & Psotka, 1994; Shute & Zapata-Rivera, 2010, 2012; Sinatra, Ososky & Sottilare, 2017; VanLehn, 2008), trends in assessment, and the opinions of the authors. Figure 1 shows a summary of the main points of the SWOT analysis.

The strengths of assessment in ITSs include:

- *Available data (response and process data)*. ITSs offer the opportunity to engage learners for long periods of time while they learn about different topics. This creates multiple opportunities to gather process and response data that can be used to infer learners' KSAs. These data can be used to refine learner models and the adaptive features of the system.
- *Various sources of evidence and levels of granularity*. Learners' data at various levels of granularity and from different sources (e.g., simulation actions, responses to dialogues, and other types activities, and data from sensors) can be used as evidence to support claims about learners' knowledge, skills, and abilities (KSAs). Macro and micro adaptive cycles can be implemented using these data (McCalla & Greer, 1994; VanLehn et al., 2007).
- *Continuous, adaptive assessment and learning loop*. ITSs implement a continuous assessment loop that aligns well with the needs for immediate adaptive feedback and other adaptive features in

learning environments. Learner models in ITSs maintain an up-to-date representation of the learner's KSAs (Shute & Zapata-Rivera, 2010, 2012; VanLehn, 2008).

- *Actionable feedback to inform instruction*. Learner model information can be made available to teachers and learners in the form of open learner model (OLM) interfaces, on-line reports, and dashboards to support learning and teaching processes (Bull, 2020; Zapata-Rivera, 2020).
- *A variety of assessment approaches including computational cognitive models, probabilistic models and machine learning*. A variety of top-down and bottom-up approaches to learner modeling have been developed and used in the field of ITSs. These approaches make it possible to manage uncertainty in ITSs (Abyaa, Idrissi & Bennani, 2019; Chrysafiadi & Virvou, 2013; Zapata-Rivera & Arslan, 2021).

The weaknesses of assessment in ITSs include:

- *Evidence framework (i.e., alignment from observables, evidence to claims)*. Assessment in ITSs could benefit from implementing an evidence framework that facilitates evidence identification and evidence aggregation processes. This evidence framework can be useful for identifying the evidence needed to support claims about learners' KSAs (Katz et al., 2017). Reusing evidence across ITS is a challenge (Robson, ITS Assessment PADLET 2021; Zapata-Rivera, et al., 2017). In addition to an evidence framework, technologies such as xAPI can be used to support this challenge (Blake-Plock, et al., 2020; Johnson et al. 2017).
- *Validity and fairness issues (e.g., accessibility).* One way of supporting the valid use of assessment information in ITSs is by improving their internal validity structure (Katz et al., 2017). Improving the structural validity of ITSs can positively impact the development of adaptive features and the appropriate use of ITSs by teachers and learners. Also, fairness issues such as improving accessibility support in ITSs should be addressed (Hansen, Zapata-Rivera & White, 2018).
- *Support for other purposes (e.g., summative).* Due in part to the goal of providing adaptive instructional support for individuals, using assessments embedded in ITSs for other purposes (e.g., certification purposes) may require additional work. This work may involve validity studies for the intended purposes, reliability, generalizability and comparability analysis.
- *Support for various stakeholders (e.g., teachers and administrators)*. Assessment information can provide useful information to other stakeholders such as teachers and administrators (Zapata-Rivera & Katz, 2014). A list of types of assessment information for different types of users can be found in Zapata-Rivera, Graesser, Kay, Hu & Ososky, S. (2020).
- *Additional support for diverse students (e.g., modeling sociocultural issues and context of learning).* ITSs could be improved by designing tasks that take into account sociocultural aspects of the learner and other aspects of the learning context. Socioculturally responsive ITSs can provide learners with more opportunities to demonstrate what they know or can do in a context that learners find engaging and at an appropriate level of challenge.

The opportunities of assessment in ITSs include:

- *Improving validity by implementing an explicit representation of evidence*. ITS platforms such as GIFT (Sottilare, et al., 2017; Johnson et al., 2017) can facilitate the implementation of an interpretation/evidence layer (Zapata-Rivera et al., 2017). Also, standardization efforts (Sottilare et al., 2018) can be instrumental in the creation of ITSs that can share assessment information across various systems in a scalable manner.

- *Implementing approaches that take into account cognitive, noncognitive, metacognitive and contextual variables.* ITSs provide multiple opportunities to model a variety of learner variables and the learner context. ITSs can use this information to refine learner models and improve on adaptive mechanisms designed to keep learners engaged and support learning.
- *Leveraging Open Learning Models (OLM) research to produce information for particular stakeholders*. As learner models become more refined using response and process data to support assessment claims, OLM interfaces, reports or dashboards can be designed and used to share learner model insights with learners, teachers, administrators and other stakeholders. These insights can facilitate learning, teaching and other decision-making processes (Bull, 2020; Kay, Zapata-Rivera, & Conati, 2021; Zapata-Rivera, 2019).
- *Improving accessibility.* Opportunities for providing support for diverse groups of learners can improve adoption of ITSs. Gathering and interpreting assessment information from these learners with disabilities may require modifications to tasks and interpretation modules to guarantee proper access and appropriate propagation of evidence (Hansen et al., 2018).

The threats of assessment in ITSs include:

- *Bias and fairness issues.* Potential bias and fairness issues with some machine learning approaches is a threat to the effective use of assessment information (Loukina, Madnani & Zechner, 2019; The Royal Society, 2019; Toreini et al., 2020). Efforts toward making learner model inferences interpretable such as human-in-the-loop approaches and mechanisms to evaluate the quality of inferences and the effects of adaptations can be instrumental in addressing these threats (Zapata-Rivera & Arslan, 2021).
- *Inappropriate use of assessment results.* Appropriate use of assessment information can contribute to supporting trust and adoption of ITSs. Learners, teachers, and other users (e.g., researchers) may be interested in knowing how the learner's assessment information is used by the system and by users to make decisions. Evaluating how users make use of assessment information provided by ITSs should be part of the development and evaluation cycle of ITS (Zapata-Rivera, 2020).
- *Security and privacy issues.* Security and privacy issues play an important role on systems that adapt their behaviors to the user. Having access to the information used by the system and how this information will be used across systems in the same ecosystem or a different ecosystem should be an important feature of adaptive learning systems (Anwar & Greer, 2012; Zapata-Rivera & Greer, 2004; Zapata-Rivera, 2020).

Future Work

Technologies and approaches explored in ITSs contribute to the development of innovative assessments. Similarly, advances in educational assessment and measurement can inform the development and evaluation of future ITSs. Results from this SWOT analysis help us identify areas for future work taking advantage of the strengths and opportunities to improve on the weaknesses and address possible threats.

As both ITSs and innovative assessments continue to explore the use of additional sources of learner data (e.g., data about cognitive, noncognitive, metacognitive aspects of the learner) and context variables to create engaging situations that can be used to support learning, issues such as evidence identification, evidence aggregation, interpretability of learner models, accessibility, privacy, and security become more relevant. Future work in these areas will support the successful implementation of assessment in ITSs. This work includes continuing to leverage results from research in the area of OLM to inform the design and evaluation of reports and dashboards that support the needs of various users of ITSs.

Recommendations for GIFT Overall

Recommendations for GIFT include:

- Providing support for the creation of learner models that make use of cognitive, noncognitive, metacognitive aspects of the learner and contextual variables.
- Leveraging OLM research to produce actionable information for different users.
- Continuing to improve validity and fairness aspects of ITSs.
- Making xAPI profile for ITS known and used beyond GIFT (Hu, ITS Assessment PADLET 2021).
- Providing guidance via information or tools on what behaviors/task-performance map onto important theoretical constructs (Graesser, ITS Assessment PADLET 2021).
- Offering assessment capabilities (e.g., a sensory module) that could be triggered on-demand and can be used to facilitate assessment implementation (Hu, ITS Assessment PADLET 2021).
- Continuing to support standardization processes and the use of assessment information across systems (Sottilare et al., 2018).

References

Abyaa, A., Idrissi, M. K., & Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. Educational Technology Research and Development, 1-39.

Anwar, M., & Greer, J. (2012a). Facilitating trust in privacy-preserving e-learning environments. *IEEE Transactions on Learning Technologies*, 5(1), 62–73. https://doi.org/10.1109/TLT.2011.23

Baker, R. (2019, March). Some Challenges for the Next 18 Years of Learning Analytics. Keynote address, 9th International Conference on Learning Analytics.

Bennett, R. E. (2018). Educational Assessment: What to Watch in a Rapidly Changing World. *Educ. Meas. Issues Pract.* 37 (4), 7–15. doi:10.1111/emip.12231A

Blake-Plock, S., Hoyt, W. Casey, C. & Zapata-Rivera, D. (2020). xAPI Data Design and the Visualization of Learning Data Considerations for a GIFT Strategy. In R. A. M., Sinatra, A. C., Graesser, X., Hu, B., Goldberg, and A. J., Hampton (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 8 – Data Visualization*. Orlando, FL: U.S. Army CCDC - Soldier Center. 163-172.

Bull, S. (2020). There are open learner models about!. *IEEE Transactions on Learning Technologies*, 13(2), 425-448.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, 40(11), 4715–4729.

Conati, C. (2009, July). Intelligent tutoring systems: New challenges and directions. *Paper presented at the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, CA

D'Mello, S. K., Gregg, J., & Southwell, R. (2020). Machine-learned computational models can enhance the study of text & discourse: A case study using eye tracking to model reading comprehension. *Discourse Processes*, 57(5–6), 420–440.

D’Mello, S. K., Tay, L., & Southwell, R. (2022). Psychological measurement in the information age: Machine-learned computational models. *Current Directions in Psychological Science*, 31(1), 76-87.

Graesser, A. (2021*). ITS Assessment PADLET Comment*. URL: https://padlet.com/xiangenhu/te59u1avsdcqq98j

Hansen, E. G., Zapata-Rivera, D., & White, J. (2018). Framework for the design of accessible intelligent tutoring systems. In S. D. Craig (Ed.). *Tutoring and Intelligent Tutoring Systems* (pp. 69-101). New York, NY: Nova Science Publishers.

Hu, X. (2021). *ITS Assessment PADLET Comment*. URL: https://padlet.com/xiangenhu/te59u1avsdcqq98j

Johnson, A., Nye. D. B., Zapata-Rivera, D.,& Hu, X. (2017). Enabling Intelligent Tutoring System Tracking with the Experience Application Programming Interface (xAPI). In R. Sottilare, A. Graesser, X. Hu, and G. Goodwin (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Assessment Methods*. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9893923-9-6. 41–45.

Kay, J., Zapata-Rivera, D., & Conati, C. (2020). The GIFT of Scrutable Learner Models: Why and How. In R. A. M., Sinatra, A. C., Graesser, X., Hu, B., Goldberg, and A. J., Hampton (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 8 – Data Visualization*. Orlando, FL: U.S. Army CCDC - Soldier Center. 25-40.

Katz, I.R., La Mar, M.M., Spain, R., Zapata-Rivera, D., Baird, J., & Greiff, S. (2017). Validity Issues and Concerns for Technology-based Performance Assessments. In R. Sottilare, A. Graesser, X. Hu, and G. Goodwin (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Assessment Methods*. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9893923-9-6. 209–224.

McCalla, G. I., & Greer, J. E. (1994). Granularity-based reasoning and belief revision in student models. In *Student modelling: The key to individualized knowledge-based instruction* (pp. 39-62). Springer, Berlin, Heidelberg.

Lehman, B., Sparks, J. R., & Zapata-Rivera, D. (2018) When should adaptive assessments care? In N. Guin & A. Kumar (Eds.), *Exploring Opportunities for Caring Assessments Workshop at the Intelligent Tutoring Systems Conference*. pp. 87–94.

Lehman, B., & Zapata-Rivera, D. (2018). Student emotions in conversation-based assessments. *IEEE Transactions on Learning Technologies*, 11(1), 1-13.

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. *In Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, (pp. 1–10), Florence, Italy.

Mislevy, R. J. (2019, October) *Current and future state of digital assessments*. Personal communication.

Mislevy R. J. (2021) Next Generation Learning and Assessment: What, Why and How. In: von Davier A.A., Mislevy R.J., Hao J. (eds) *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment*. Methodology of Educational Measurement and Assessment. Springer, Cham. https://doi.org/10.1007/978-3-030-74394-9_2

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

Mislevy, R. J., & Yan, D. (2017). Evidence-Centered Assessment Design and Probability-Based Inference to Support the Generalized Intelligent Framework for Tutoring (GIFT). In R. Sottilare, A. Graesser, X. Hu, and G. Goodwin (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Assessment Methods*. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9893923-9-6. 101-123.

Robson, R. (2021). *ITS Assessment PADLET Comment.* URL: https://padlet.com/xiangenhu/te59u1avsdcqq98j

Shute, V. (2019, October) *Current and future state of digital assessments*. Personal communication.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for a learning system called ACED. *International Journal of Artificial Intelligence in Education*, 18(4), 289–316.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessmen*t, 21(1), 34-59.

Shute, V. J., & Psotka, J. (1994). Intelligent Tutoring Systems: Past, Present, and Future. Armstrong Lab Brooks AFB TX Human Resources Directorate.

Shute, V. J. & Zapata-Rivera, D. (2010). Educational measurement and Intelligent Systems. In E. Baker, B. McGaw, & P. Peterson (Eds.), *Third Edition of the International Encyclopedia of Education*. Oxford, UK: Elsevier Publishers. vol. 4, pp. 75-80.

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach, & A. Lesgold (Eds.), *Adaptive technologies for training and education*. New York, NY: Cambridge University Press. 7-27.

Sinatra, A. M., Ososky, S., & Sottilare, R. (2017). Assessment in Intelligent Tutoring Systems in Traditional, Mixed Mode, and Online Courses. In R. Sottilare, A. Graesser, X. Hu, and G. Goodwin (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Assessment Methods*. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9893923-9-6. 235-247.

Sottilare, R., Barr, A., Robson, R., Hu, X., & Graesser, A. (2018). Exploring the opportunities and benefits of standards for adaptive instructional systems (AISs). In Proceedings of the Adaptive Instructional Systems *Workshop in the Industry Track of the 14th International Intelligent Tutoring Systems* (pp. 49-53).

Sottilare, R., Graesser, A.C., Hu, X., & Goodwin, G. (Eds.) (2017). Design Recommendations for Intelligent Tutoring Systems: Assessment (Vol. 5). Orlando, FL: U.S. Army Research Laboratory.

Sparks, J.R., Peters, S., Steinberg, J., James, K., Lehman, B.A., & Zapata-Rivera, D. (2019). Individual Difference Measures that Predict Performance on Conversation-Based Assessments of Science Inquiry Skills. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Sparks J.R., Zapata-Rivera D., Lehman B., James K., & Steinberg J. (2018). Simulated Dialogues with Virtual Agents: Effects of Agent Features in Conversation-Based Assessments. In: Penstein Rosé C. et al. (eds) *Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science*, vol 10948. Springer, Cham, 469-474.

The Royal Society (2019). Explainable AI: the basics policy briefing. Available at https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.

Toreini, E., Aitken, M., Coopamootoo, K., Elliot, K., Gonzalez-Zelaya, C., and van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Conference on Fairness, Accountability, and Transparency* (FAT'20), Barcelona, Spain.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. T*he future of assessment: Shaping teaching and learning*, 113-138.

VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., & Rose, C.P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31,, 3-62.

Verschelden, C. (2017). *Bandwidth Recovery: Helping students reclaim cognitive resources lost to poverty, racism, and social marginalization*. Sterling, VA: Stylus.

Zapata-Rivera, D. (2017). Toward Caring Assessment Systems. *In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17*). ACM, New York, NY, USA, 97-100. DOI: https://doi.org/10.1145/3099023.3099106.

Zapata-Rivera D. (2019). Supporting Human Inspection of Adaptive Instructional Systems. In: Sottilare R., Schwarz J. (eds) Adaptive Instructional Systems. *HCII 2019. Lecture Notes in Computer Science*, vol 11597. Springer, Cham. pp. 482-490.

Zapata-Rivera, D. (2020). Open Student Modeling Research and its Connections to Educational Assessment. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-020-00206-2

Zapata-Rivera D., & Arslan, B. (2021). Enhancing Personalization by Integrating Top-down and Bottom-up Approaches to Learner Modeling. In: Sottilare R., Schwarz J. (eds) Adaptive Instructional Systems. *Adaptation Strategies and Methods. HCII 2021. Lecture Notes in Computer Science*, vol 12793. Springer, Cham. pp. 234-246.

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, and G. Goodwin (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 5 - Assessment Methods*. Orlando, FL: U.S. Army Research Laboratory. ISBN 978-0-9893923-9-6. 125–136.

Zapata-Rivera, D., Graesser, A., Kay, J., Hu, X., & Ososky, S. (2020). Visualization Implications for the Validity of ITS. In R. A. M., Sinatra, A. C., Graesser, X., Hu, B., Goldberg, and A. J., Hampton (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 8 – Data Visualization*. Orlando, FL: U.S. Army CCDC - Soldier Center. 61-68.

Zapata-Rivera, J. D., & Greer, J. (2004). Inspectable Bayesian student modelling servers in multi-agent tutoring systems. *International Journal of Human Computer Studies*. 61(4), 535-563

Zapata-Rivera, D., & Katz, R. I. (2014): Keeping your audience in mind: applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice.* 21(3), 442-463.

Zapata-Rivera, D., Lehman, B., & Sparks, J.R. (2020). Learner Modeling in the Context of Caring Assessments. In: Sottilare R., Schwarz J. (eds) *Adaptive Instructional Systems. HCII 2020. Lecture Notes in Computer Science*, vol 12214. Springer, Cham. pp. 422-431.

Chapter 8 – Team Tutoring in Intelligent Tutoring Systems SWOT Analysis

Peter W. Foltz[1] and Stephen B. Gilbert[2]
University of Colorado, Boulder[1]; Iowa State University[2]

Introduction

A good team can accomplish more than the efforts of its individual members. However, individual members need to have the skills to perform as a team. Skills such as interpersonal understanding, proactivity, decision-making as a group, and agreement in roles all increase effectiveness of group performance (Druskat & Kayes, 2000; Prichard et al., 2006). These skills are also recognized as being important for hiring decisions. In 2011, 90 executives surveyed counted teamwork in the top 10 soft skills desired in employees (Robles, 2012). Despite this need, the field's understanding of the theoretical basis of teamwork is still evolving, and there are not systematic methods for assessing and teaching team skills (Graesser et al., 2020; Lai, DiCerbo & Foltz, 2017). Thus, while Intelligent Tutoring Systems (ITSs) have been successful in offering automated learning scaffolding for individual students, Intelligent Team Tutoring Systems (ITTSs), which offer software-based coaching to a team of humans, have yet to mature fully. This is due to a variety of factors including the complexity of capturing team data, limited techniques for analyzing team performance, and the need for effective skills frameworks and methods for displaying key team metrics.

A key challenge for implementing ITTSs is the lack of formalized methods and technical approaches for measuring the quality of teamwork. On a team there are task skills (how well the team performs at a specific task) and team skills (qualities of the team that might transfer to other tasks). While team researchers have developed frameworks of skills (e.g., Hesse et al., 2015; OECD, 2017a), proposed critical factors in successful teaming (e.g., Salas et al., 2018), and have even proposed measurable behavioral markers for some of these factors (Rosen & Foltz, 2014; Sottilare et al., 2018), robust computational models of team dynamics are still in their infancy (e.g., Gorman et al., 2017). Concurrently, while we must consider how humans work together, as software agents and robots become more sophisticated, the research area of Human-Autonomy Teaming (HAT) has become an important related area that asks similar questions: How can we measure the effectiveness of the working relationship between Astronaut X and Agent A? How can we assess whether Soldier Y works better with Robot 1 or Robot 2? If we can establish metrics for these constructs, then we could theoretically build systems that train HATs and human teams to perform better together.

Whether the team is composed of a group of learners with one tutor or a tutor for each learner, an effective ITTS that could develop team skills as well as task skills while focusing on each team member's individual needs could be invaluable. This SWOT analysis describes the current state of the art of ITTS development and the pathways to the future.

SWOT Analysis

The following SWOT analysis was conducted by drawing on the individual research findings of the authors, a review of the state of the research literature, investigation of existing ITTSs, as well as on the community of researchers who have assembled this book. The goal of the analysis was not to develop a comprehensive list of all possible SWOT elements, but to focus on the key elements that are involved in advancing the field and then organize those elements into groups of major factors that influence the development and success of work in the field.

Strengths

The primary strength of ITTSs stems from their potentially broad applicability to a variety of domains. So many of today's business tasks are performed by teams that team skills are needed by a majority of workers. For example, the 2016 Job Outlook Survey of the National Association for Colleges and Employers report that nearly 80 percent of respondents look for evidence that candidates can work in a team (National Association for Colleges and Employers, 2016). However, another survey (Casner-Lotto & Barrington, 2006) showed that only 25 percent of employers characterized four-year college graduates' teamwork skills as "excellent" and other work has shown that university faculty are not well prepared to teach team skills and do not value it highly (Chen et al., 2004). Thus, another strength is that ITTSs can fill a pressing need in the workforce.

If an ITTS could teach team skills effectively, those skills would add value across multiple team tasks. Ideally, a set of reusable modules for teaching team skills could be created that would use the domain of the task as input (e.g., a communication module could be directed to give practice tasks from the pilot domain or surgery domain, and it would measure a team's communication performance as well as give them feedback on that performance). For this cross-task generalization approach to work, we would need a common infrastructure that could support a variety of team tasks. These components of an ITTS should be agnostic to whether the team members are human or software agents, so that the system could assess team dynamics independently of the composition of the team.

Weaknesses

Much of the difficulty of team tutoring systems can be categorized as "resolving ambiguity" when measuring what the team is doing (Sottilare, Team Tutoring Systems PADLET 2021). For example, measuring communication of a team can be difficult because people may speak at the same time, which leaves an automated transcription system uncertain as to the current speaker. Also, even if communication can be accurately logged, as linguists know, analyzing individual words can be futile without understanding and classifying the larger context of the communication. Next, many team behaviors are non-verbal, e.g., nods, crossing of arms, pointing, etc., which makes them difficult to record without many cameras and a good computer vision body recognition system, so that the system records not only the nod from Team Member A, but also the response to the nod from Team Member B.

This mix of verbal data, non-verbal data, and task performance data leads to mathematical models that feature rich multimodal interactions, which can be very valuable, but because of complexity and the number of teamwork variables, these models require a significant amount of data to be made robust. Typically, team studies are expensive to run and coordinate, but there is potential hope from YouTube, in which numerous multiplayer video gaming teams have logged their performance. For models of multiplayer game team behavior, at least, there may be plenty of data available.

Another challenge in developing ITTSs is the creation and distribution of appropriate feedback for the teams and team members (e.g., Stevens et al., 2019). This challenge has several facets. First, the ITTS will likely assess the team at both a team level (How is the team doing?) and an individual level (How is Maria doing? How is Giovanni doing? etc.). If the ITTS is sophisticated enough to provide general feedback based on both levels, it is a non-trivial challenge to decide how much of that feedback to give without overwhelming the members. If the team tutor has two suggestions for the team overall and three critiques for Giovanni, should Giovanni receive all five pieces of feedback, or is that too much? While expert human coaches may have an intuitive understanding of how to balance these, even potentially personalizing their decision based on Giovanni's personality and how well he responds to feedback, it is difficult to make this expertise explicit and embody it in code. Even if there were good authoring tools for ITTSs, often the team coach has domain expertise but not the technical expertise to author the different cases or conditions in which each set of feedback would be offered.

Opportunities

As the field of ITTSs is still very new, there are a wide range of opportunities for advancing both research and development of effective training systems. The largest opportunity that exists is that there is great demand for training of teams. As the world has moved from an industrial society to an information society, there is increased need for supporting collaboration in the complex socio-technological systems that require people to work together effectively (e.g., Autor et al., 2003). However, the workforce is not yet prepared. The PISA (Program for International Student Assessment) assessment showed that only 8% of 15-year-olds across OECD (Organisation for Economic Co-operation and Development) member countries performed at the highest level of proficiency in collaborative problem solving (OECD 2017b), and managers overwhelmingly report that graduating college students do not have the requisite skills for collaboration (American Management Association, 2012). Thus, a tutor that could train team skills for a range of domains would provide enormous return on investment (ROI) if proven effective. Also, given the rise in global connectivity and telecommuting, ITTSs could be beneficial in smoothing teamwork between members from different global cultures.

To achieve such a generalized ITTS, there are opportunities to improve our understanding of the nature of teamwork and collaboration. This work can include the development of frameworks of skills and mapping how feedback can be applied to improve specific skills. Studies of effective and ineffective teams can provide empirical bases of when and how feedback can be most useful. Additionally, more research can be done on improving our approaches to measurement of skills. This work can include determining what kinds of overt and covert behaviors are effective indicators of performance and how to apply better models for measuring the complex multi-modal, multi-person dynamics. Such approaches can benefit from contributions from a variety of fields, including psychometrics, team science, dynamic modeling, and educational training in order to understand the effects of tasks and contexts and how they interact with the measurement of team skills.

A final opportunity revolves around applying novel technology to improve ITTSs. Team tasks generate a wealth of data including speech, logs of actions, facial expressions, gestures, typed communication, and task-related outputs. These multi-modal interactions provide multiple pieces of converging evidence about individual and team states. Recent developments in Artificial Intelligence (AI) technology can provide approaches to reduce the data into features related to team skills and develop models that can predict team cognitive, social and affective states (e.g., Butler & Randall, 2013; Calvo & D'Mello, 2010; Richardson et al., 2007) . These approaches include automated speech recognition, deep-learning models of language and visual features such as facial expressions and gestures, reinforcement learning and dynamic models of changes in states, and machine learning-based methods for combining multimodal data (Grafsgaard et al., 2014; Vinciarelli & Esposito, 2017). Overall, these approaches provide great opportunities for broadening the kinds of team interaction data that can be captured and modeled in order to provide more effective feedback.

Threats

Developing tutors for individuals is hard, requiring effective models of a person's learning state, techniques for tracking learner understanding, and approaches for measuring gain. Yet researchers have been highly successful in creating tutors for individuals. Developing tutors for teams adds another level of challenges on top of all the methods needed for individual assessment. For good team situations that can be tutored, there needs to be a balance of open, realistic interactions, while also controlling the training situation well enough to assess performance. The key challenges that need to be overcome in order to bring about successful team tutoring systems are addressing the variability in types of team tasks, the complexity of the team and individual assessment methods, and techniques for collecting multimodal, multi-party data.

Team tasks can vary significantly from each other. Teams can have many varied structures, with different roles, leadership configurations, and characteristics of team tasks such as shared or different goals (Bonner et al., 2015). For example, some team tasks may have strong interdependence, like a car assembly line team, while some have lower levels, like a team of call center customer support agents. Yet, many of the skills such as shared understanding, achieving common purpose and maintaining a team organization are still required, but may vary based on the organizational structures and task characteristics. This level of variability makes it more difficult to have more generalizable methods for training.

The level of complexity of team tasks further provides challenges for developing training situations and assessment methods. Training often focuses on task skills and tasks can be simulated effectively with synthetic task environments and computer-based simulations. However, due to the dynamics of multiple people interacting, it can be more difficult to create circumstances which simulate potential team situations that would be useful for training. For example, to create a situation where a team member fails to come to consensus with the rest of the team may require either an actor or an intelligent agent to play the role. Use of actors can be expensive, while use of intelligent agents requires development of effective AI that can understand the training context and have the agent intervene in ways that elicit the targeted skills and respond appropriately to the learners (e.g., Bergner et al.., 2016; Rosen & Foltz, 2014). The complexity in the team tasks also requires methods that can track the individual actions and team interactions and process the stream of information in a way to assess the desired team skills, to determine whether team members have achieved appropriate levels of those skills, and when and how the skills should be remediated. To achieve this requires additional research and development of frameworks of team skills incorporating social, cognitive and affective states, continued work on AI based assessment approaches, and psychometric methods to measure those skills. With such complex assessment models, it becomes more difficult to update them for new tasks, adapt them for new situations, as well as extract the psychometric features in a way that can make the judgments made by the assessment models explainable in a way that they can be turned into effective feedback and training for the team. Indeed, just because we can assess a team skill does not mean that we know how and when to intervene to remediate.

Finally, we should recognize that data in team tasks is hard to collect. Much of the interaction data may be verbal, but some information may be conveyed through other modalities including hand and facial gestures and sharing of task artifacts. This can require instrumenting for the collection of audio, video and task log files, annotating with behavioral codes, and then processing the data to obtain a complete picture of the state of the team. Synthetic environments can allow more control over data collection. However, in realistic environments, such as classrooms, battlefields, and Combat Information Centers, instrumenting the data collection can be logistically difficult, requiring collaboration and integration with the ongoing processes and systems of the learning environment.

Overall SWOT Analysis

Below we provide the overall SWOT analysis generated during the workshop which summarizes the above narrative.

Strengths of ITTSs

Scoring/Feedback available in more structured environments and domains.
Common infrastructure means you can build for a variety of team tasks.
Expanding set of reusable modules allows use of new input types and better modeling of learning progressions.
Enable more practice of higher-level thinking skills.
Could tutor on both task skills and team skills.
Could make predictions of whether Learner A will work well with Learner B on future team working on a different task: cross-task generalization.
Multi-modal interactions give multiple pieces of converging evidence (speech, language, affect, actions, etc.).
Most work is done by teams, so tutors play towards skills needed by a majority of workers.

Weaknesses of ITTSs

Methods for assessing performance, scoring, and feedback in complex and/or multi-person simulations are not mature.
Level of technical expertise/need for domain experts for building training insufficient.
Currently we do not have software architecture to support one tutor knowing about both individuals and the team. Instead, it’s a team tutor co-present with a tutor for each individual (“crowd of tutors”), and the tutors cannot share knowledge with each other. Learners ask, ”Which tutor should I listen to?”
Need a “hierarchical” tutor framework.
Difficult to measure some team dynamics and interactions (e.g., non-verbal communication).
Many contextual factors to control for if you want to do effective modeling of real-world situations.
Literature on when to give feedback to individuals vs. whole team is mixed.
Need complicated models to model rich multimodal interactions.
What is the role of the tutor for a team? Guide, monitor, facilitator, shepherd?
Ability to resolve ambiguity during team training and educational events is limited.

Opportunities for ITTSs

Dynamic models of learning and collaboration for tracking performance could be quite valuable.
Automated content and assessment item creation for teams would be useful.
Continued development of improved approaches for automated scoring of open-ended responses/team language would make significant contributions to the field of communication research.
Human-in-the-loop creation/deployment of intelligent training systems would aid numerous contexts.
AI-based explainability approaches can add transparency, accountability and interpretability of outputs from AI-based assessment models.
Agent-based approaches, part of Human-Autonomy Teaming, are in high demand.

A good team tutor domain module for team skills would be applicable for almost any team task and provide an enormous ROI if done well.
There are many athletic coaches who do this well and who could be observed/interviewed as models for when to give whole-team feedback and when to give individuals feedback.
Team Tutor studies generate an enormous amount of data to be analyzed.
Zoom Cloud Recording has great transcription and video quality, which can aid team analysis.

Threats

Challenges of collecting good, clean, exploitable data in realistic training environments are significant.
Team tasks vary significantly (e.g., tasks with strong interdependence, like a car assembly line vs. tasks with very little, like a team of call center customer support agents).
Potential user/public backlash against using AI for assessing higher level thinking skills.
More complex assessment models lead to greater difficulty in updating, adapting, and providing explainability in training modules.
Teams are complicated, with many different structures (some with different roles, some with same role, different leadership configurations, etc.) This complexity threatens generalization possibilities and makes this work expensive.
Just because we know how to assess it, doesn't mean we can teach it.
If student data is used to train an AI model, gaining appropriate permissions can be non-trivial.

Supporting Research

Research in Collaborative Problem Solving, Team Assessment and Team Training has been growing over the last decade as demand for more effective methods and training technologies has increased. Below, we highlight several areas of research that are supporting some of the primary areas of growth in this field.

One key area is in *defining skills and how to measure them*. This work has included the development of a variety of frameworks which break down collaboration and teamwork skills, operationally define the skills, and describe approaches to measuring them (see Hesse, et al., 2015; von Davier et al., 2017; O'Neil et al., 1995; OECD, 2017a; Sottilare et al., 2018; Sun et al., 2020). These theories of measurement have focused on how to align skills constructs to evidence of behavior (e.g., Hao et al., 2019) and adopting an evidence-centered design approach to analyzing team skills (e.g., Mislevy, Steinberg & Almond, 1999).

Measurement of teams has also been approached by applying *AI-based methods to analyze team data*. Teams produce a wealth of data while participants interact, and a key challenge is to reduce the data in ways to find evidence of skills. AI-based methods are ideally suited for this approach since they can learn important patterns of team interactions and sift through data efficiently to detect emergent behaviors that are indicative of effective or ineffective team performance. Because much of the data generated by teams is verbal communication, advances in natural language processing have shown great promise for converting language streams into performance data (e.g., Cooke et al., 2012; Foltz & Martin, 2009; Foltz et al., 2006). Because team performance is not generally characterized as a skill occurring at any particular instance, but instead is the result of the processes that occur over time, process data is critical to track and monitor these changes. Thus, techniques to analyze the interactional dynamics based on log data are of particular note (e.g., Dunbar, et al., 2020; Morgan et al., 2012).

Thirdly, the field of team science is realizing that team data is generated from many sources and therefore, *multimodal, multiparty techniques* are required to account for the rich data that occurs during any team collaboration. This work can include looking at measuring factors such as speech, body movement, eye gaze, predictions of affect, and log data. Techniques can then be used to combine these measurements to paint a fuller picture of the interactions and cognitive, social and affective states of the participants (e.g., Calvo & D'Mello, 2010; Eloy et al., 2019).

A final relevant growth area of note is Human-Autonomy Teaming (HAT). Other related terms are human-agent teaming, human-robot interaction (HRI), and human automation interaction, but human-autonomy teaming seems to be the emergent term (Lyons et al., 2021). HATs are teams consisting of humans and one or more agent, but the agents are typically teammates rather than tutors or coaches as in an ITTS. All the same, research focused on analyzing HATs overlaps with some of the frameworks described above, characterizing members of HATs in terms of their interdependence, their communication, their authority relationships, and other ITTS-relevant factors (Ouverson et al., 2018; Sepich et al., 2021).

Recommendations for GIFT in Particular

Based on the findings above, and previous experience building ITTSs with GIFT (Gilbert et al., 2017; Gilbert et al., 2018), several recommendations for GIFT and other ITTS support systems emerge. First, while GIFT currently supports team tutoring via a crowd of tutors (one tutor for the team and one tutor for each team member), there is not currently any prioritization mechanism for balancing the influence of each tutor, e.g., deciding when the team member should receive feedback from their individual tutor vs. the team tutor, and how to ensure that the amount of feedback given is appropriate to the context. Second, team tutor authoring is burdened with the particular difficulty of thorough testing, with its potentially exponential number of training system states (e.g., based on a series of decisions by different team members). Automation tools that aid in this quality assurance testing will be critical. Lastly, templates for authoring ITTSs for different contexts would be very helpful, e.g., for an ITTS for group problem solving, decision making, highly interdependent tasks vs. low ones, tasks under time pressure vs. not, etc. Having a set of ITTS templates based on a sensible taxonomy of team structures and team tasks would save ITTS authors significant time.

Discussion and Recommendations for Future Research

In summary, there is great promise in research, development, and deployment of Intelligent Training Systems for Teams. The SWOT analysis permits us to identify and focus on the key elements needed for progress in the field. Overall, the key strengths include the ability to model team dynamics and the characteristics of team members that contribute to those dynamics. The key weaknesses include the logistical difficulty of measuring subtle team interactions and the lack of precise knowledge about how to best balance the mix of feedback to the whole team vs. to individual members. Key opportunities include the ability to tutor on team skills that span multiple task contexts and the prevalence of human coaches as role models. Key threats include the overall complexity and variance within team tasks, which affect both data collection for training ITTS and its generalizability beyond a single context. Overall, the results of the SWOT analysis show that ITTSs are difficult to build but have great potential and are well worth pursuing.

To realize the potential of ITTSs will require advances in science and technology as well as changes in educational focus and policies. Curricular reform, which puts more emphasis on CPS skills, is growing with educational standards starting to incorporate required collaboration skills (e.g. Fiore et al., 2017; National Research Council, 2011; OECD, 2017b). These standards will flow into classrooms to ensure that students are trained with the requisite skills. Concurrently, industry must put more emphasis on these skills as requirements for graduates and/or adopt team training as part of continuing education.

Advances in research will require continued recognition of team assessment as inherently multidisciplinary and entail incorporating advances from multiple fields to move the field forward. While much of the prior focus of research has been on assessing cognitive skills in teams, integrating the social and affective states of team members will be critical. The most promising areas for research include: psychometric measurement for complex performance, natural language processing, multimodal fusion, affect detection, and dynamical time series analyses. With advances in computer-based agents, more work will also be needed in human-agent teams to understand how best to apply agents, both as team-members that can supplement human performance but also as trainers who can improve how humans perform as team members. Put together, these approaches can also help develop a taxonomy of the types of teams and tasks that are most appropriate for ITTSs and pathways to building effective training systems.

References

American Management Association . (2012). Executive summary: AMA critical skills survey. Workers need higher level skills to succeed in the 21st century. Retrieved from http://www.amanet.org/uploaded/2012-Critical-Skills-Survey.pdf

Autor, D. H., Levy, F., Murnane, R. J., The Skill Content of Recent Technological Change: An Empirical Exploration, The Quarterly Journal of Economics, Volume 118, Issue 4, November 2003, Pp. 1279–1333, https://doi.org/10.1162/003355303322552801

Bergner, Y., Andrews, J. J., Zhu, M., & Gonzales, J. (2016).Agent-based modeling of collaborative problem solving(Research Report No.RR-16-27). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.1211

Bonner, D., Gilbert, S., Dorneich, M. C., Burke, S., Walton, J., Ray, C., & Winer, E. (2015). Taxonomy of Teams, Team Tasks, and Tutors. In Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2) (pp. 189-198).

Butler, E.A. and A.K. Randall, Emotional coregulation in close relationships. Emotion Review, 2013. 5(2): p. 202-210.

Calvo, R. A. and D'Mello, S. (2010) Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 18-37, doi: 10.1109/T-AFFC.2010.1.

Casner-Lotto, J., & Barrington, L. (2006). Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce. Washington, DC: Partnership for 21st Century Skills.

Chen, G., Donahue, L. M., & Klimoski, R. J. (2004). Training undergraduates to work in organizational teams. Academy of Management Learning and Education, 3(1), 27–40.

Cooke, N. J., Duchon, A., Gorman, J. C., Keyton, J., Miller, A. Preface to the Special Section on Methods for the Analysis of Communication. Human Factors. 2012;54(4):485-488. doi:10.1177/0018720812448673

Druskat, V. U., & Kayes, D. C. (2000). Learning versus performance in short-term project teams. Small Group Research, 31(3), 328–353.

Dunbar, T. A., Gorman, J. C., Grimm, D. A., & Werner, A. (2020). The Dynamical Systems Approach to Team Cognition: Theories, Models, and Metrics. In Contemporary Research(pp. 53-74). CRC Press.

Eloy, L., EB Stewart, A., Jean Amon, M., Reinhardt, C., Michaels, A., Sun, C., ... & D'Mello, S. (2019, October). Modeling team-level multimodal dynamics during multiparty collaboration. In 2019 International Conference on Multimodal Interaction (pp. 244-258).

Fiore, S. M. et al. Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress (National Center for Educational Statistics, United States Department of Education, Washington DC, 2017).

Foltz, P. W., & Martin, M. J. (2009). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches (pp. 411–431). Routledge/Taylor & Francis Group.

Foltz, P. W., Martin, M. J., Abdelali, A., Rosenstein, M. B., & Oberbreckling, R. J. (2006, July). Automated team discourse modeling: Test of performance and generalization. In Proceedings of the 28th Annual Cognitive Science Conference(. pp. 1317-1322).

Gilbert, S. B., Dorneich, M. C., Walton, J., & Winer, E. (2018). Five Lenses on Team Tutor Challenges: A Multidisciplinary Approach. In J. Johnston, R. Sottilare, A. M. Sinatra, & C. S. Burke (Eds.), Building Intelligent Tutoring Systems for Teams (pp. 247-277). Emerald Publishing. https://doi.org/10.1108/S1534-085620180000019014

Gilbert, S., Slavina, A., Sinatra, A. M., Bonner, D., Johnston, J., Holub, J., MacAllister, A., Dorneich, M., & Winer, E. (2017). Creating a Team Tutor Using GIFT. International Journal of Artificial Intelligence in Education, 28, 286-313. https://doi.org/10.1007/s40593-017-0151-2

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.

Grafsgaard, J.F., J.B. Wiggins, A.K. Vail, K.E. Boyer, E.N. Wiebe, and J.C. Lester. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. 2014.

Gorman J. C., Dunbar T. A., Grimm D., Gipson C. L. (2017). Understanding and Modeling Teams As Dynamical Systems. Frontiers in Psychology. Jul 11;8:1053. doi: 10.3389/fpsyg.2017.01053. PMID: 28744231; PMCID: PMC5504185.

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. ETS Research Report Series, 2019(1), 1-17.

Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In Griffin, P., Care, E. (Eds.), Assessment and teaching of 21st century skills: Methods and approach (pp. 37–56). Dordrecht, The Netherlands: Springer.

Lai, E. R., DiCerbo, K. E., Foltz, P. (2017). Skills for today: What we know about teaching and assessing collaboration. Retrieved from http://www.p21.org/storage/documents/Skills_For_Today_Series-Pearson/Collaboration_White_Paper_FINAL.pdf

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J. D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Handbook of research on technology tools for real-world skill development (pp. 344-359). IGI Global.

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–Autonomy Teaming: Definitions, Debates, and Directions. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.589585

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (1999). Evidence-centered assessment design. Princeton, NJ: Educational Testing Service.

Morgan, B., Keshtkar, F., Duan, Y., & Graesser, A. C. (2012). Using state transition networks to analyze multi-party conversations in a serious game. In. S. A. Cerri, & B. Clancey (Eds.), Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012) (162-167). Berlin: Springer-Verlag.

National Association of Colleges and Employers. (2016). Job outlook 2016. Bethlehem, PA: National Association of Colleges and Employers.

National Research Council (2011). Assessing 21st century skills. Washington, DC: National Academies Press.

OECD (2017a). PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving (Rev. ed.), Retrieved from the OECD website: http://dx.doi.org/10.1787/9789264281820-en

OECD (2017b). PISA 2015 Results (Volume V): Collaborative Problem Solving, PISA. Paris: OECD Publishing.

O'Neil, H. F., Chung, G. K., & Brown, R. S. (1995). Measurement of teamwork processes using computer simulation. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Graduate School of Education & Information Studies, University of California, Los Angeles.

Ouverson, K., Iglesias Pena, M., Walton, J., Gilbert, S., & Dorneich, M. (2018). What Intelligent Team Tutoring Systems Can Learn from Human-Agent Teams. In Proceedings of Technology, Mind & Society (pp. Article 28).

Prichard, J. S., Stratford, R. J., & Bizo, L. A. (2006). Team-skills training enhances collaborative learning. Learning and Instruction, 16(3), 256–265.

Richardson, D.C., R. Dale, and N.Z. Kirkham, The art of conversation is coordination common ground and the coupling of eye movements during dialogue. Psychological Science, 2007. 18(5): p. 407-413.

Robles, M. M. (2012). Executive perceptions of the top 10 soft skills needed in today's workplace. Business communication quarterly, 75(4), 453-465.

Rosen, Y. & Foltz, P. W. (2014). Assessing collaborative problem solving through automated technologies. Research & Practice in Technology Enhanced Learning, 9(3).

Salas, E., Reyes, D. L., & McDaniel, S. H. (2018). The science of teamwork: Progress, reflections, and the road ahead. American Psychologist, 73(4), 593.

Sepich, N., Dorneich, M. C., & Gilbert, S. B. (2021). Human-Agent Team Game Analysis Framework: Case Studies. In Proceedings of the Human Factors & Ergonomics Society (HFES) Annual Meeting (Vol. 65, pp. 1146-1150).

Sun, Chen, Shute, Valerie J., Stewart, Angela, Yonehiro, Jade, Duran, Nicholas, & D'Mello, Sidney. (2020). Towards a generalized competency model of collaborative problem solving. Computers & Education, 143 ©. https://doi.org/10.1016/j.compedu.2019.103672

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J., & Gilbert, S. B. (2018). Designing Adaptive Instruction for Teams: A Meta-Analysis. International Journal of Artificial Intelligence in Education, 28, 225-264. https://doi.org/10.1007/s40593-017-0146-z

Sottilare, R. (2021). Team Tutoring Systems PADLET Comment.

Stevens, R., Gorman, J., Zachary, W., Johnston, J., Dorneich, M., & Foltz, P. How is this team doing and why? Design recommendations for intelligent tutoring systems, 77.

Vinciarelli, A. and A. Esposito, Multimodal Analysis of Social Signals, in Handbook of Multimodal-Multisensor Interfaces. 2017.

von Davier, A. A., Hao, J., Liu, L., & Kyllonen, P. (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: Lessons learned from developing a collaborative science assessment prototype. Computers in Human Behavior, 76, 631-640.

Chapter 9 – Self-improving Systems in Intelligent Tutoring Systems SWOT Analysis

Min Chi[1], Xiangen Hu[2], and Gautam Biswas[3]
North Carolina State University[1]; University of Memphis[2]; Vanderbilt University[3]

Overview: What is a Self-Improving Learning System/Environment

Self-Improving systems can be loosely referred to as systems that have the capacity to monitor, evaluate and improve their own performance as a function of experience. Such systems need to learn from experience, therefore, they must include machine learning, or some other data-driven technique that supports learning and changes in behavior over time. In this chapter, we consider a minimalist view of *human learning systems* (such as brick-and-mortar schools, or virtual learning systems). From this perspective, *learning systems consist of* four interconnected components: *Human Learners, Learning Resources, Learning Processes, and Learning Environments* (Kuo & Hu, 2019). While learning systems are improving as a whole (they have improved over the years due to advances in the learning sciences and education theories, technology, and policy), all of the four components are also *improving* in their own ways: *human learners* improve when learning happens; *learning processes* improve when better theories of learning are implemented; finally advances in technology make learning environments and learning resources improve when appropriate technologies are applied.

We consider a system *"self" improving* when it is improving without the "*explicit*" help of other systems. It is obvious that human learners have self-improving capabilities. Some of the learning resources (such as teachers, tutors, and human study mates) also have self-improving capabilities. Only recently, several types of *dynamic* digital learning resources, such as intelligent tutoring systems (ITSs), have shown promise that certain learning resources may also have self-improving capabilities (Wenger, 2014). This chapter focuses on the self-improving capabilities of these types of learning resources. We can examine self-improving capabilities at a system level such as learning eco-systems (that contain all four components), or at the level of individual components. For the purpose of this chapter, we consider the "*self*-improving" capabilities of each of the four components. Specifically, we examine the self-improving capabilities of ITSs in the context of the Generalized Intelligent Framework for Tutoring (GIFT, Sottilare et al., 2012).

The remainder of this chapter is organized as follows: we will first describe the three aspects of self-improving learning systems: *self-improvable, self-improvability, and self-improving*; then we will give three examples of ITS implementations that had certain levels of self-improving capabilities; to conclude, we will offer a Strengths Weaknesses Opportunities and Threats (SWOT) analysis of self-improving systems within GIFT and provide recommendation and future research.

Three Aspects of Self-Improving Learning Systems: Self-Improvable, Self-Improvability, and Self-Improving

There are three closely related yet different aspects of self-improving learning systems: *Self-improvable, Self-improvability, and Self-improving*. Next, we will describe them in detail.

Self-Improvable systems can change their behaviors based on their interactions with learners, and such changes are driven not by hard-coded rules but by data that is gained from experiences when operating in an environment. To make it self-improvable, a system should be designed with the following necessary properties*:*

1. Existence of “master” memory (Data Store) that captures system behavior and experiences. The memory includes the interaction history of similar systems and learners.
2. Existence of variable controllable components. All the variables can be changed at run-time.
3. Application programming interfaces (APIs) that connect variable controllable components with the Data Store.
4. A collection of ideal (effective and efficient) instructional strategies to guide the use of the APIs.

The first three properties make the systems changeable, while the last property makes the system changeable in a way that improves student learning.

Self-Improvability is *the degree* that a self-improving system improves. Inspired by the development of self-driving vehicles, it is useful to consider *levels* of self-improving capabilities prior to the existence of real self-improving learning systems**.** Table 1 shows a total of six “Levels” of self-improvability of self-improving learning systems that we have defined, from minimum self-improvability (level 0) to complete self-improvability (level 5).

Table 1: Six Levels of Self-improvability

Levels	Descriptions
0	Observe the learner’s behavior (passive) and select (mechanically) a pedagogy from a pre-specified list. For example, system behavior is independent of the learner’s recent interactions with the system.
1	The selection of the pedagogy is a function of the learner’s interaction history.
2	Prior to the selection of pedagogy, the system classifies students’ behavior based on the cognitive nature of the task.
3	Prior to selection of pedagogy, the system maps students’ behavior based on the ***cognitive and non-cognitive*** nature of the task (to a pre-specified list). The system also observes and evaluates the outcome of the selected pedagogy.
4	The system dynamically learns how to classify a student's behavior (cognitive / non-cognitive), based on previous scenarios. The system also observes and evaluate the outcome of selected pedagogy.
5	The system creates and validates new pedagogy that was never observed or used previously. It observes, evaluates the outcomes, and confirms created pedagogy.

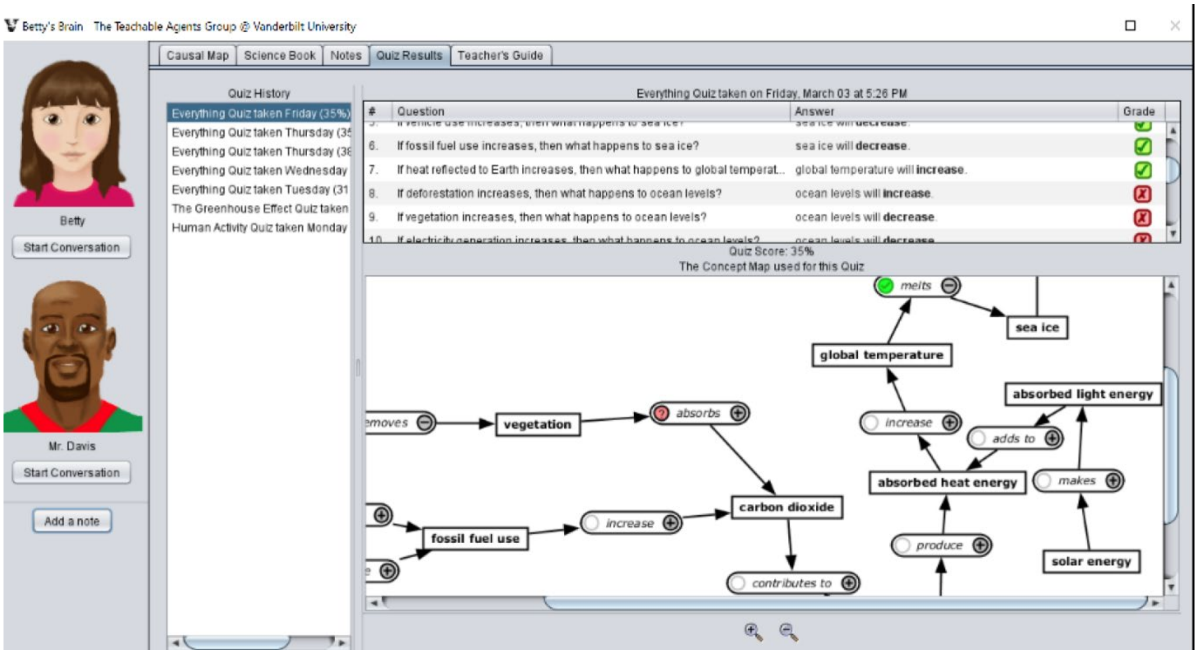
Self-improving is the process that makes a self-improvable learning system achieve levels of self-improvability. The process includes 1) selecting/creating technology that enables appropriate input from the learner to the systems; 2) selecting/creating an appropriate data structure that stores system and learner behavior; 3) selecting/creating appropriate assessment models for both system and learner behavior; 4) selecting/creating appropriate APIs that connect system components to the assessment outcome (of the system and learner) and alter parameters of systems components; and 5) evaluating the outcome of 1-4 to make the learning system meet the desired level of self-improvability.

Current State-of-Art Most on Self-Improvable

In this section, we review a few examples of ITSs that meet some levels of self-improvability.

Betty's Brain [Level 2 – 3]

The Betty's Brain system is designed to make science learning an active, constructive, and engaging process for students (Biswas et al., 2005; 2016). A primary innovation in this intelligent and adaptive open-ended learning environment is that it leverages the learning by teaching paradigm (Bargh & Schul, 1980; Biswas et al., 2005) to get students to research and construct models of science phenomena in the guise of teaching a virtual agent generically called Betty. Students actively engage with Betty during the learning process by building a causal model of a scientific process, asking her questions, and getting her to take quizzes that are provided by a mentor agent named Mr. Davis. Figure 1 illustrates the quiz interface for the Betty's Brain system.



Everything Quiz taken on Friday, March 03 at 5:26 PM

#	Question	Answer	Grade
5.	If vehicle use increases, then what happens to sea ice?	sea ice will **decrease**.	✓
6.	If fossil fuel use increases, then what happens to sea ice?	sea ice will **decrease**.	✓
7.	If heat reflected to Earth increases, then what happens to global temperat...	global temperature will **increase**.	✓
8.	If deforestation increases, then what happens to ocean levels?	ocean levels will **increase**.	✗
9.	If vegetation increases, then what happens to ocean levels?	ocean levels will **decrease**.	✗
10.	If electricity generation increases, then what happens to ocean levels?	ocean levels will **decrease**.	✗

Quiz Score: 35%

The Concept Map used for this Quiz

Figure 1. Quiz Interface of the Betty's Brain System

When asked to answer questions or to take a quiz, Betty uses a qualitative reasoning mechanism to chain together a sequence of links and generate answers like, "*If deforestation increases, the amount of heat trapped by the earth will increase*" (Leelawong & Biswas 2008). Betty's performance on the quizzes provides students with the feedback that students needed to check their map and come up with strategies for identifying and correcting errors and omissions in their maps. When asked, Betty can provide explanations for her derived answers, and this helps students identify and analyze the individual links that she uses to generate her answers. Betty also provides motivational feedback by expressing happiness when her scores on the quiz improve, and she expresses disappointment when her quiz scores do not improve. Additional feedback is provided by the mentor agent in the form of learning strategies that students can employ when they are not performing well.

In recent work, Munshi et al. (2022) have improved the adaptability of the Betty's Brain system by *tracking* the learners' performance on building the causal map, their learning strategies (e.g., their Read → Map Building, or Quiz → Map Building strategies), as well as their affective state (e.g., delight, confusion, and frustration) and providing *tailored feedback* through Mr. Davis and Betty to help students improve their learning and map building performance as they work on the system. Machine learning algorithms, such as sequence mining (Zaki, 2001) and differential sequence mining (Kinnebrew et al., 2013) supported by analytics methods, such as coherence analysis (Segedy et al, 2015) form the basis for the adaptability in the Betty's Brain system that drives the feedback system to support learners Kinnebrew et al., 2017). Therefore, the system exhibits Level 2 – 3 self-improving characteristics by adapting to the learners needs. Overall, in classroom studies, Betty's Brain has been very effective in helping students develop metacognitive strategies to become better learners, learn about causal models of scientific processes (e.g., pond ecology, climate change, and human body thermoregulation), and apply these models to problem-solving tasks.

AutoTutor [Level 2 – 3]

AutoTutor is an ITS that engages in natural language conversations with the learner, as described in research by Graesser et al. (1999) and Nye et al. (2014). Its effectiveness has been demonstrated in various fields, including computer literacy, physics, and critical thinking. The system's design incorporates three key research areas: human-inspired tutoring strategies, pedagogical agents, and technology that enables natural language tutoring ("AutoTutor," n.d.). For the purpose of this chapter, we highlight several components of AutoTutor that make it a self-improving adaptive instructional system (SIAIS); further details can be found in Hu et al. (2019).

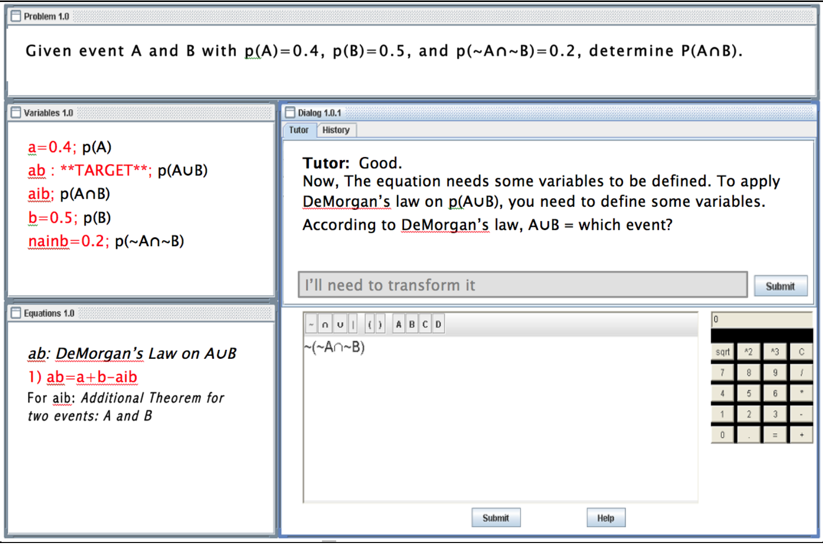
Expectation-misconception tailored (EMT) dialog: The central feature of AutoTutor is its dialog management, known as EMT dialog. The EMT dialog works by evaluating the learner's answers and providing hints and prompts based on whether the answer matches the expected answer keys or exhibits common misconceptions. This approach ensures that the learner receives personalized instruction and is guided towards a deeper understanding of the material.

Avatar: AutoTutor uses conversational avatars to interact with learners, creating an engaging learning experience. The avatars provide hints and prompts in natural language and have unique personalities that can be controlled parametrically. These avatars can exhibit different attitudes and emotions based on the learner's behavior, which can motivate and engage learners. The ability to control the personality of the avatars using data makes AutoTutor an SIAIS.

AutoTutor Scripts: AutoTutor Scripts are XML files that contain all dialog moves and parameters that guide AutoTutor's behavior. These scripts also include expected right answers and typical misconceptions for a given question, as well as semantic answers and regular expressions for these expected answers and misconceptions. By including this information in the scripts, AutoTutor is able to provide personalized instruction that is tailored to the learner's needs and misconceptions. The use of XML ensures that the content is structured and easily accessible, making it an effective tool for instruction. One key feature (that is relevant to the current chapter) of AutoTutor Scripts is that they include parameters such as thresholds for matching the answers and misconceptions. These parameters can be dynamically changed during runtime, allowing AutoTutor to self-improvable based on its interactions with learners. This adaptability ensures that AutoTutor is responsive to the needs of the learner, and can continue to improve its instruction over time.

AutoTutor Rules: AutoTutor Scripts contain a collection of "if-then" statements known as rules, which are a key component that make AutoTutor a SIAIS (Semi-Intelligent Automated Instructional System). These rules use the learner's coverage of expectations or exhibition of misconceptions as the condition (if), and the corresponding hint or prompt as the action (then). These functions are parametrically configured and can be altered at runtime, making AutoTutor highly adaptable to the learner's needs.

Pyrenees Probability Tutor: A Self-improving Learning System Through Deep Reinforcement Learning (DRL) [Level 2].



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

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

ab: DeMorgan's Law on A∪B
1) ab=a+b-aib
For aib: *Additional Theorem for two events: A and B*

Tutor: Good.
Now, The equation needs some variables to be defined. To apply DeMorgan's law on p(A∪B), you need to define some variables.
According to DeMorgan's law, A∪B = which event?

I'll need to transform it

~(~A∩~B)

Figure 2. Pyrenees probability tutor showing the problem, variables, equations, dialog, and student workspace

Pyrenees (Figure 2) is an ITS that teaches the introduction of probability and conditional probability (Zhou, Azizsoltani et al., 2021; Zhou, Azizsoltani et al., 2019; Zhou, Yang, et al., 2019). *One prominent self-improvability of Pyrenees is through a Reinforcement Learning-induced pedagogical policy.* To make Pyrenees self-improving over time, in partnership with North Carolina State University faculty since 2014, exploratory training corpora and experimental data have been collected by training undergraduates on Pyrenees. In the exploratory mode, the systems are constrained to make random yet reasonable pedagogical decisions. This exploratory corpus approach has also been widely used in previous research on the application of reinforcement learning (RL) to improve dialogue systems (Williams et al., 2005). For each semester, we apply various reinforcement learning approaches to derive pedagogical policies with the goal to improve student learning gains. The existing pedagogical agent's policy is simultaneously updated in Pyrenees. The updated system is then used to interact with students with some baseline models (such as Pyrenees following the best policy from previous semesters). The effectiveness of the induced policies is empirically evaluated, and the newly collected student-system interaction logs are added to the training corpus for the next round of pedagogical policy induction. To date, our exploratory and experimental corpora for Pyrenees includes more than 2000 student-system training interaction logs. Next, we will describe our approach in detail.

For many learning environments, the system-learner interactions can be viewed as a sequential decision process and RL offers one of the most promising approaches to data-driven pedagogical decision-making for improving student learning. Figure 3 illustrates how pedagogical policy induction can be represented as traditional RL. At any given time *t*, the pedagogical *agent* observes the environment state *s* (a vector representation of relevant learning context features as shown in Table 2 below), then chooses an action *a*, and receives a reward *r* (calculated from success measures), and the environment transitions into state *s'*. The agent learns the policy by estimating the action-value function *Q(s,a),* defined by the following Bellman Equation:

$$Q(s,a) = R(s,a) + \Sigma_{s',t'} \gamma P(s,a) max_{a' \in A} Q\left(s',a'\right),$$

where *R(s,a)* is the immediate reward, γ is a discount factor, and we sum the discounted Q-values of the optimal action a' for each possible next state s' (using transition probabilities p(s'|s, a), estimated from the training corpus). Once the optimal action-value function $Q^*$is found, the optimal policy is to take the action with the highest Q-value.

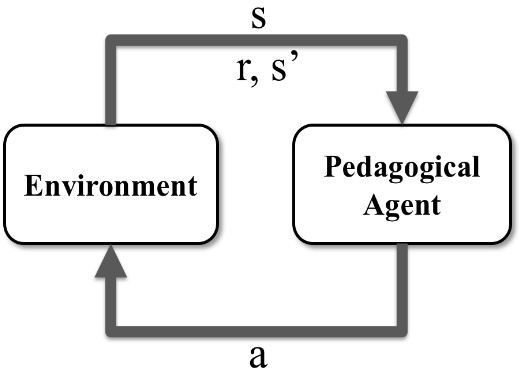


Figure 3. Pedagogical policy induction

For RL, as with all machine learning, success depends upon having an effective state representation *S* to model the environment. We used more than 140 state features including both cognitive and non-cognitive features such as the student's current knowledge level, affective state, the task, and other salient features suggested by literature (D'Mello & Graesser, 2010; Koedinger & Aleven, 2007). Table 2 below shows five categories of state feature that describe student learning process and the system behaviors.

Table 2. Selected Feature Examples

Feature Family	Selected Example Features
Student Engagement	Time since the last action.
	The number of Worked Examples the student has received since the last Problem solving.
	Total number of Worked Examples or Problem solving on the current problem and over the whole tutor.
Learning Context	Difficulty level of a problem.
	The number of different types of applied rules for current problems.
Student Performance	Number and percentage of the correct student problem solving steps.
	The number of subgoals achieved on the current problem.
Student Actions	Total number of hints/skips that students have taken so far.
	The total number of worked examples or problem solving adopted into problems / ignored.
Temporal Situation	Time since the tutor last provided an intervention.
	Elapsed time on the current session.
Student Strategies	Presence/absence of temporal subgoal achievement patterns.

One primary challenge for RL pedagogical policy induction is delayed rewards. Just as supervised learning models depend heavily on accurate output labels, RL approaches depend heavily on an accurate reward function. Generally, immediate rewards are more effective than delayed rewards because the more delay, the harder it becomes to assign credit or blame properly. However, the most appropriate ITS reward is learning gains, which are unavailable until the training is complete. This is due to the complex nature of the learning process which makes it difficult to assess students' learning *moment by moment* and more importantly, many instructional interventions that boost short-term performance may not be effective long-term. We applied two different approaches: a Gaussian Processes (GP) based approach (Azizsoltani et al., 2019) and a general deep neural network approach to *infer* immediate rewards from delayed rewards (Ausin et al., 2021). Results from a series of experiments showed that using inferred immediate rewards can indeed lead to better RL policies than using delayed rewards.

The combination of deep learning and novel reinforcement learning algorithms has made solving complex problems possible with deep reinforcement learning (DRL) (Andrychowicz et al., 2020; Silver et al., 2018; Vinyals et al., 2019). Despite DRL's great success, there are many challenges in order for DRL to be applied successfully to ITSs. One major challenge is sample inefficiency: DRL algorithms often need millions of interactions to learn a policy. There exists, however, a sub-field of RL, named batch RL (also known as offline RL) which can learn the optimal policy from a small amount of data to generalize to unseen scenarios (Lange et al., 2012) as our recent work showed that a policy-induced by applying batch DRL with GP-based inferred rewards is significantly more effective than an expert-designed policy (Ausin et al., 2019). We further combined deep reinforcement learning and hierarchical RL to induce hierarchical DRL policies that would decide whether to offer worked examples or problem-solving at two *fixed* granularities: problem and step levels (Zhou, Azizsoltani et al., 2019).

The ITSs' Six Levels of Self-improvability

In Table 3, below, we will describe the six levels of self-improvable capabilities ITSs can provide by describing the input data captured and focusing on two critical components of the ITS that can leverage the data: learner models and pedagogical models.

Table 3. Levels of Self-improvability

For the Level 0 systems, we have:

	Descriptions of Level 0 System
Input Data	Learner's current problem-solving actions.
Learner Models:	Learner models are based only on the observations of current categorical behavior, such as correct or wrong responses of simple assessment items.
Feedback /Pedagogical	Immediate. Correct/Incorrect.

For the Level 1 systems, we have:

Level 1	**Descriptions of Level 1 systems**
Input Data	Tracking students' activities over time.
Learner Models:	Learner models are based on activities of current AND immediate past (current learning episode), including non-categorical data such as response latency.
Feedback /Pedagogical	Immediate after error. But based on aggregated learner model for current learner episode.

For the Level 2 systems, we have:

Level 2	**Descriptions of Level 2 systems**
Input Data	Tracking students' activities over time.
Learner Models:	Learner models aggregate information over multiple learning episodes of students. They are based on activities of current AND immediate past (current and previous learning episode) to capture learner behaviors and strategies, as well as non-categorical data such as response latency and bio metrics. Cognitive and non-cognitive factors are considered, such as motivation, affect, and learning environments.
Feedback /Pedagogical	Learners are allowed to explore. Feedback only after repeated instances of the same error. Feedback provided is over students' aggregated performance overall learning episodes.

For the Level 3 systems, we have:

Level 3	**Descriptions of Level 3 systems**
Input Data	Tracking students' activities using log files + speech modalities.
Learner Models:	Modeling students' metacognitive behaviors that combine cognitive and metacognitive processes. Learner model tracks evolutions of learners' behaviors over time.
Feedback /Pedagogical	Go beyond performance to also study students' learning behaviors, i.e., their sequence of activities over time and their related performance. Feedback at strategic behavior level.

For the Level 4 systems, we have:

Level 4	**Descriptions of Level 4 systems**
Input Data	Tracking students' activities using multiple modalities: log files + speech + gesture+ video + eye tracking.
Learner Models:	Modeling students' Cognitive, Affective, Metacognitive, and Motivational (CAMM) processes along with their interactions with the environment, for example, with their instructors, and other students.
Feedback /Pedagogical	Extend feedback mechanisms to account for students' Self-regulated learning (SRL) processes and to help them better interact with the environment.

For the Level 5 systems, we have:

Level 5	**Descriptions of Level 5 systems**
Input Data	Tracking students' collaborative (teamwork) processes.
Learner Models:	Distributed cognition. Study students' learning in the context of the space they are learning in. Computer learning environment + other learners (collaborative learning) + interacting with artifacts in the room to learn.
Feedback /Pedagogical	Extend feedback mechanisms to account for students' SRL processes in an online fashion.

Self-improvable, Self-improvability, and Self-improving of the Six Levels of ITSs

Next, we will summarize how each level of ITSs map to these three aspects: *Self-improvable, Self-improvability, and Self-improving*. To make it *self-improvable*, a system should have parametrically controllable components, a master memory (data), and an API connecting the data with parametric control components guided by learning outcomes (so that they can be improved). *Self-improvability* is the degree that a self-improving system improves and *self-improving* is the process that makes a self-improvable learning system achieve higher levels of self-improvability. Table 4 shows a mapping of the six levels to the three aspects (self-improvable, self-improvability, and self-improving).

Table 4. Mapping the Six Levels to the Three Aspects

Level	Self-Improvable (offline)	Self-Improvability (measures)	Self-Improving (online/offline learning)
0	No	Static	No
1	No	adaptive-- reactive (locally)	No
2	Partially	adaptive-- temporal	No
3	Yes	adaptive -- multimodal (strategic behaviors)	No
4	Yes	adaptive – SRL, shared regulation (CAMM processes)	No
5	Full	Spatio-temporal learning environments	Yes

SWOT Analysis of Self-improving ITSs

Advancement of theories and technologies made it possible for researchers to build Self-improving ITSs at the first three levels (0,1, and 2), as seen in the examples above. Thus, we will focus on providing SWOT analysis for the three highest levels: levels 3-5.

SWOT for Level 3:

- The ***strength*** of current technologies, especially high fidelity and speed of data collection, large capacity of data storage, and advanced big-data processing technique make it possible for recording students' learning activity.
- The ***weakness*** is the lack of processes that effectively utilize the large volumes of learner data.
- The weakness offers the ***opportunities*** for learning scientists to make use of the data to propose and validate computationally feasible models and processes.
- ***Threats:*** Having computationally feasible models and processes from data is the key for Self-improving ITSs. Without this, we will only have self-improvable systems that only are self-improving up to lower-level self-improvability.

SWOT for Level 4:

- ***Strength:*** some of the learning systems (such as GIFT, with the sensor module) have shown promises that the modern technology can capture and record learners' behavior beyond log files.

Any computer systems currently being sold can accept multimedia inputs. This is a strength and sufficient condition for building the level for Self-improving ITSs.

- The ***weakness*** here provides opportunities for learning scientists and engineers to develop breakthrough technologies and theories with possible new research areas. It is a playground for learning sciences and AI researchers.
- ***Opportunities:*** when dealing with multimedia input (voice, emotion, etc.), that data is usually "noisy". In addition, the model and/or process are not sensitive enough to establish clear relations between observed multimedia input (during learning) to learners' Cognitive, Affective, Metacognitive, and Motivational (CAMM) processes.
- ***Threats:*** When considering multimedia input, such as speech, facial emotion, etc., there is a danger of privacy and security, with potential biases. It is possible that multimedia data collected are biased. For example, the Self-improving ITSs may build a learner model based only on part of the learner's data.

SWOT for Level 5:

- ***Strength:*** Social psychology provides theories on individual and team behavior in general. Existing systems started to consider learning in collaborative environments. These are strengths for level 5 Self-improving ITSs.
- ***Weakness:*** There is a lack of computationally feasible model and process to allow any Self-improving ITSs to successfully consider learners' data in collaborative learning environments.
- ***Opportunities:*** similar opportunities as the previous levels.
- ***Threats:*** similar Threats as the previous levels.

Self-Improvement Framework and Existing Systems

In this section, we revisit our descriptions of the three systems presented earlier: Betty's Brain, AutoTutor, and Pyrenees, and discuss their current self-improvability status plus what may be needed to migrate them to higher level self-improving systems. We characterized all three of these systems as Level 3 on the self-improvability scale, which implies they are adaptive to learners' performance and the cognitive and metacognitive strategies they employ to support their learning and problem solving processes. However, these systems have not reached the level of understanding to support students' self-regulation processes that focus on learners' abilities to understand and control their learning behaviors, i.e., their cognitive, affective, metacognitive, and motivational processes, which help them to accomplish their learning and problem-solving goals (Azevedo et al, 2017; Panadero, 2017). Self-regulation emphasizes the students' autonomy, strategy use, self-monitoring, and self-reflection during problem-solving, all necessary traits for being successful life-long learners (Winne & Hadwin, 1998; Zimmerman, 2002).

Similarly, these systems are not designed to operate effectively in collaborative learning scenarios, where students learn to solve problems in small groups, interact with other objects in their environment (e.g., additional tools and experimental setups that support learning), and with humans (e.g., their instructors) to aid their learning and problem solving processes. Analyzing such distributed cognition scenarios, requires efficient collection and analysis of multimodal (e.g., vision, speech, eye tracking, and system log) data, which is currently not well-integrated into current computer-based intelligent tutoring systems (ITSs). Therefore, our current systems are still not at Level 4 of the self-improvability scale.

Last, these systems use advanced machine learning techniques, such as sequence mining, and RL to support pattern detection and decision making, and to improve system design to make them more adaptive. However, these methods are currently implemented offline, with humans analyzing the results generated by the machine learning algorithms to improve system performance and adaptability. Since learning does not currently happen online, these systems are yet to reach the levels of autonomy and decision making that would help them to achieve Level 5 of the self-improvability scale.

References

Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A. & Schneider, J. (2018). Learning dexterous in-hand manipulation. arXiv preprint arXiv:1808.00177.

Ausin, M. S., Azizsoltani, H., Barnes, T., & Chi, M. Leveraging Deep Reinforcement Learning for Pedagogical Policy Induction in an Intelligent Tutoring System. In Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019) (Vol. 168, p. 177).

Ausin, M. S., Azizsoltani, H., Ju, S., Kim, Y.-J., & Chi, M. (2021). InferNet for Delayed Reinforcement Tasks: Addressing the Temporal Credit Assignment Problem. 2021 IEEE International Conference on Big Data (Big Data 2021), Orlando, FL, USA, December 15-18, 2021, 1337–1348.

Azizsoltani, H., Kim, Y. J., Ausin, M. S., Barnes, T., & Chi, M. (2019). Unobserved is not equal to non-existent: using Gaussian processes to infer immediate rewards across contexts. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 1974-1980). AAAI Press.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. Journal of Educational Psychology, 72(5), 593-604.

Biswas, G., Segedy, J.R., & Bunchongchit, K. (2016). From Design to Implementation to Practice – A Learning by Teaching System: Betty's Brain. International Journal of Artificial Intelligence in Education, 26(1), 350-364.

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005). Learning by teaching: A new agent paradigm for educational software. Applied Artificial Intelligence, 19, 363-392.

D'Mello, S. K., & Graesser, A. C. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction, 20(2), 147–187.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. Cognitive Systems Research, 1(1), 35–51.

Hu, X., Cai, Z., Graesser, A. C., & Cockroft, J. L. (2019, May). GIFT as a Framework for Self-Improvable Digital Resources in SIAIS. In *Proceedings of the 7th Annual GIFT Users Symposium* (p. 49). US Army Combat Capabilities Development Command–Soldier Center.

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2017). Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. IEEE Transactions on Learning Technologies, 10(2), 140-153.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. Journal of Educational Data Mining, 5(1), 190-219.

Koedinger, K. R., & Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. Educational Psychology Review, 19(3), 239–264.

Kuo, B.-C., & Hu, X. (2019). Intelligent learning environments. Educational Psychology Review, 39(10), 1195–1198.

Lange, S., Gabel, T., & Riedmiller, M. (2012). Batch reinforcement learning. In Reinforcement learning (pp. 45-73). Springer, Berlin, Heidelberg.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. International Journal of Artificial Intelligence in Education, 18(3), 181-208.

Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., & Paquette, L. (2022). Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. Journal of Computer Assisted Learning. https://doi.org/10.1111/jcal.12761

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. International Journal of Artificial Intelligence in Education, 24(4), 427–469.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. & Lillicrap, T. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140-1144.

Sinatra, A. M. (2018, May). Team models in the generalized intelligent framework for tutoring: 2018 update. In Proceedings of the Sixth Annual GIFT Users Symposium (Vol. 6, p. 169). US Army Research Laboratory.

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

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

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P. & Oh, J. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 1-5.

Wenger, E. (2014). Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge. Morgan Kaufmann.

Williams, J. D., Poupart, P., & Young, S. J. (2005). Factored partially observable Markov decision processes for dialogue management. In Int. Joint Conf. on Artificial Intelligence (pp. 393–422).

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. Machine learning, 42(1), 31-60.

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.

Zhou, G., Yang, X., & Chi, M. Big, Little, or Both? Exploring the Impact of Granularity on Learning for Students with Different Incoming Competence. (2019). In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), Proceedings of the 41st Annual Conference of the Cognitive Science Society (pp. 3206-3212). Montreal, QB: Cognitive Science Society.

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2021). Leveraging Granularity: Hierarchical Reinforcement Learning for Pedagogical Policy Induction. International Journal of Artificial Intelligence in Education, 1–47.

Chapter 10 - Data Visualization in Intelligent Tutoring Systems SWOT Analysis

Judy Kay[1], Arthur C. Graesser[2], and Anne M. Sinatra[3]
University of Sydney[1]; University of Memphis[2]; U.S. Army Combat Capabilities Development Command (DEVCOM) Soldier Center[3]

Introduction

Intelligent Tutoring Systems (ITSs) can be designed to collect rich learning data. This chapter considers how visualizations of that data can be made useful for key stakeholders in educational settings. There are many potential stakeholders, including learners, their peers, teachers, parents, people who manage the funding, those reporting the effectiveness, system developers and researchers in education and educational technology. There are also many contexts for education and the way that a learning system fits into them. The Generalized Intelligent Framework for Tutoring (GIFT) is designed to try to accommodate all stakeholders by emphasizing reusability and domain independent authoring tools. . While all these stakeholders and contexts are important, one key area that data visualization in GIFT needs to formally address is learners and instructors in formal settings, which is the focus of this chapter.

In the decades of research on ITSs and Artificial Intelligence in Education (AIED), learning data was designed so that the system could build and maintain a learner model; that learner model is essential as it drives the personalized teaching and learning that are defining features of ITSs and AIED. A central question is how to also make use of that learning data, and the learner model, to create visualizations that help both students and instructors advance their learning and instructional objectives. One important body of AIED research that has tackled this challenge calls such a visualization an *Open Learner Model (OLM)* (Bull, 2020; Bull & Kay, 2016) - this is because it *opened* the system's model of the learner. A more recent field of research, called learning analytics, has a slightly different emphasis. It explores how to capture learning data and to create interfaces onto that data for the teachers/instructors (Bodily et al., 2018). Both approaches require the design of both *learner-facing dashboards* and *instructor-facing dashboards*. A key difference is that the design of the OLM is based on a learner model rather than learning data. Across all this research, there has been a bewildering diversity of visualizations and terms to describe them. As we identify the strengths, weaknesses, opportunities, and threats for ITS data visualization, we introduce definitions and frameworks to underpin understanding of the diverse previous work.

A valuable foundation for the design of any visualization is to identify the key stakeholders and their needs that determine the purpose of visualizations. For learners, learning data visualizations can serve the follow purposes (Bull & Kay, 2016):

- support metacognitive processes of planning, monitoring and reflection;
- share and discuss data with peers to collaborate or to compete;
- make choices about navigating through the system;
- check the data, correct errors in the system, potentially including data that is for another person;
- exercise the right to access and control learners' personal data.

The design of visualizations that address these needs can be guided by identifying *system benchmark questions.* These are questions that learners should be able to answer from the visualization. Table 1 gives examples of these, as articulated in Kay and Kummerfeld (2019) and in Kay et al. (2022), which introduced the term benchmark question since this follows best practice in designing and evaluating interfaces in terms of the tasks that can be used in evaluation studies (Hartson & Pyla, 2012). See Table 1 for examples of benchmark questions.

Table 1. Benchmark Questions

ID	Abstract Benchmark Question	Examples of concrete questions
1	Am I making progress?	In this problem solving activity, was my last step making progress?
2	Am I meeting my own goals, over the short or long term?	Have I reached mastery level in this topic? Have I reached mastery in all the topics?
3	Am I meeting external goals	Am I on track for a pass? Am I behind the rest of the class? Am I behind the top 10% of the class?
4	What changes might help me reach my goals?	Should I stop working on this problem and reread the teaching materials?
5	What is the meaning of the data and components modeled?	The system models my reading skills - what does that mean?
6	Can I trust the accuracy of the model?	A friend did the last problem for me - so does this model show my knowledge?

The first three questions are core for informing learning. To answer these questions, a student needs a visualization of their own data in a suitable form. For Question 1, a student needs to check their progress as they complete each learning activity. For Question 2, the information needs to be available in a form that enables the student to judge recent progress against their target.

For Question 3, the visualization needs to provide more than just the individual student's data. It should show that personal data matches *external benchmarks*. These may be mastery standards, such as the level of performance required to pass a subject. They may also be *normative* data that summarizes performance or other characteristics of fellow students. Importantly, the visualization should make it easy for the learner to compare their own data against the benchmark.

Instructors need two forms of visualizations: individual and aggregate. The first, individual form is important for many teaching roles. For example, the teacher may meet the student to discuss their learning progress, help them overcome problems, and/or create a study plan. In this level, the purposes and questions map to the student ones above. At the aggregate level, instructors need a visualization of learning data for a group of students. This is essential for teachers to self-monitor their pedagogical activities to improve their teaching and the students' learning. It is also valuable for classroom orchestration and to enable a teacher to decide which student most needs their attention.

It is important to take a *user* perspective in designing each of the learning data visualizations intended to address the needs of one or more of the stakeholders. A user perspective should account for the broader aspects of human-system engineering for functionality, usability and desirability (Roscoe et al., 2018). Figure 1 distinguishes four forms of information that visualizations may have available. The top green box is for learning data about a single student, for example their success on a problem-solving activity. The next green box is for aggregate level, such as the success for all students on that activity.

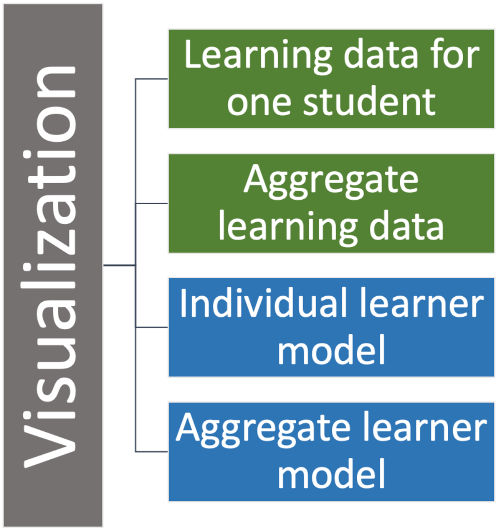


Figure 1. Each visualization may present a user with a view of learning data (green) or learner models (blue), each at the level of an individual student and aggregated data from multiple students.

The lower boxes in Figure 1 are for two forms of learner models. Much ITS work built the individual learner model that captures the system's beliefs about the learner. For ITSs, the most important components in the learner model represent the learner's knowledge. For example, a system that teaches programming in Python may have a learner model representing the user's mastery of each of the main topics taught, such as variables and loops. It may also model other aspects needed to drive personalization of the teaching, such as the student's misconceptions, preferences, goals, engagement, and psychological attributes.

Figure 1 highlights the distinction between *learning data* and a *learner model*. Learning data is any data that a system collects about the learner as part of their learning activities and assessments in a learning system. The data can be very rich and detailed. For example, a system could easily capture time-stamped records of all clicks. By contrast, a learner model is designed in three stages:

- define an *ontology* of the components to be modeled,
- mechanisms to collect just the *relevant learning data* for those components and
- a method to use that data to *infer the learner's mastery* of knowledge components and the *values* of other components relevant to learning.

The learner model ontology provides a foundation for structuring an ITS visualization in terms of the learning goals that the system was created to teach. This opens the possibility for the visualization to serve as a form of communication from the designer of the learning system to the learner as well as the instructor.

Figure 1 distinguishes two forms of learner models, namely those that model an individual learner and those that model multiple learners. The aggregate models may range from small groups, such as a single class to tens of thousands of MOOC (Massively Open On-line Course) users. Unfortunately, the ITS literature too often uses the term learner model for both of these.

The field of educational data mining has produced a body of work on learner models, often built using sophisticated machine learning. Here the term learner model typically refers to an aggregate learner model. For example, a simple learner model can be built from the data of thousands of users of a math teaching system that has open-ended problems. Then suppose an individual student has taken three steps through a solution to solve a problem. The teaching system can compare this with the aggregate learner model and inform the student about the percentage of students who correctly solved the problem after taking each of these three steps. If 80% of students who took the first two steps solved the problem but no student who took the third step did so, sharing this information could be a prompt for the student to reconsider the third step. Broadly educational data mining models are valuable for systems to make predictions about an individual, by making use of both the individual and aggregate learner models (Pelánek, 2017).

Contemporary perspectives in the learning sciences (e.g., the second volume of *How People Learn* of the National Academy of Sciences, Engineering, and Medicine, 2018) have emphasized the idiosyncratic needs of individual learners, as opposed to one-size-fits-all models. This requires learner models for individuals. The learner models are very different for individuals versus multiple learners. For the first two classes of benchmark questions in Table 1, designers need to draw on these theoretical foundations. For the third class of benchmark question, where the individual aims to make sense of their own learning data and learner model in relation to standards, educational theory also provides valuable guidance for supporting self-regulated learning.

SWOT Analysis

The experts at the GIFT workshop identified a number of strengths, weaknesses, opportunities, and threats in the SWOT analysis on data visualization. Judy Kay gave a presentation on data visualization at the meeting, followed by a discussion among the experts and comments entered in an on-line PADLET software facility. This section presents the highlights of the SWOT analysis.

Strengths

Advances in data visualization have benefitted from prior progress in relevant fields of research, and applications. These efforts were acknowledged at the GIFT workshop, as summarized below.

(1) Decades of relevant research in broader areas of cognitive science, human-computer interaction, human factors, learning technologies, data science, and information visualization.

(2) Research advances in learner modeling.

(3) Advances in open learner models (data for the learner) and digital teaching platforms (data for the instructor).

(4) Explorations of effective methods for organizing and depicting data for multiple purposes, such as for instructors (orchestration, whole class monitoring, reflection on overall learning) and for students (learning support, self-regulated learning, peer discussion).

Weaknesses

The experts identified a number of weaknesses in data visualization for ITSs and adaptive learning environments (ALEs). Part of this can be attributed to the complexity, grain-size, and diversity of measures collected in these technologies, at least compared to conventional computer-based training that typically collects a handful of measures on learning activities, such as overall performance scores, progress on a small set of topics, attendance/dropout, training time, etc. Another challenge is that learning materials in ITS/ALE technologies are highly adaptive to the learner and therefore presented under a complex set of conditions (as opposed to a rigidly scripted set of materials). This presents complications in interpreting data because there are likely selection biases when presenting a particular pedagogical activity to a particular class of students. The biases are hopefully carefully systematically addressed, but there are risks that need to be tracked and evaluated. For example, if a high performing learner receives more difficult pedagogical activities, how can the student's performance be compared to a lower performing student? Are particular learner populations put at a disadvantage when adaptive learning environments are delivered? With respect to this chapter, this presents added challenges in the design of data visualization. The weaknesses identified at the GIFT workshop are presented below.

(1) Disagreement among researchers on the terms to use and their meanings, such as the examples of learner model presented above.

(2) Minimal consensus on the definitions of different categories of learning environments and their associated methodologies for data analysis and visualization.

(3) Barriers to communication and cross fertilization among multiple related research communities, such as AIED, ITS, LAK, L@S, HCI, human factors, Infoviz, etc.

(4) Most people outside the ITS, AIED, and ALE community are not aware of or do not understand our work.

(5) Challenges for stakeholders (students, teachers, researchers, public) to understand a visualization, such as data and visual literacy, accessibility, and cognitive bias.

(6) Stakeholders (students, teachers, researchers, public) are not adequately trained on how to interpret visualizations that depict uncertainty, a key challenge for learning contexts where data is typically incomplete, uncertain and noisy, meaning that the results include considerable uncertainty.

(7) Stakeholders (students, teachers, researchers, public) may misuse visualizations of learning data.

(8) Lack of principles and guidelines for addressing very different contexts, notably the very different demands where visualization is for fast (at-a-glance) use versus slow (reflective) use (Kay et al., 2020).

(9) Limited understanding of good ways to visualize long term data.

(10) Challenges in accounting for uncertainty in learning data and honestly communicating data limitations.

(11) Failure in designing learning software for easy collection of relevant data that is aligned with meaningful data information visualization.

(12) Lack of established methods for bringing learners into a meaningful role in the entire process of collecting and interpreting their learning data, thereby giving learners agency, control and responsibility.

(13) Lack of standards for what to present and corresponding lack of stakeholder awareness of what can be useful.

Opportunities

Correcting the weaknesses is, of course, an important class of opportunities in the future of ITS/ALE environments. Experts at the GIFT workshop added a number of other opportunities for future research and development. These addressed formulating standards for data visualization design and increasing communication with researchers in other fields who can help us or benefit from our help to promote standards.

(1) Overcoming the weaknesses identified above.

(2) Timely definitions of alternative learning technologies, conceptual ideas, and ontologies with some modicum of consensus. GIFT and the IEEE learning standards movement are exemplars of such efforts.

(3) Purpose-driven data design (designing data so that it will be useful).

(4) Sharing standards with many relevant research communities that develop learning technologies.

(5) Building bridges with HCI and Infoviz communities, with a focus on the particular nature of learning data and the purposes that matter for designing learning data visualizations.

(6) Building bridges with the AI and FATE (fairness, accountability, transparency, and ethics in AI) communities to link learning and ethical concerns.

Threats

The threats articulated by Judy Kay and the experts at the GIFT workshop were very diverse and addressed multiple stakeholders in building, testing, and implementing ITS/ALE systems. A systems perspective is needed to meaningfully respond to the threats and may take decades to mitigate because some of the threats involve different generations of stakeholders who have different expectations on learning as well as digital technologies.

(1) Inappropriate use of data that conflicts with rights to privacy promotes a surveillance culture.

(2) Failure to account for all the uncertainty in learning data that may mislead stakeholders.

(3) Failure to carefully account for the gap between the intended purpose of the data visualization and what has actually been possible to create.

(4) Potential tensions between recommendations of educational theory and stakeholders who want social comparison information with risks of ignoring the theory.

(5) Failure to create systems that nurture learner control and responsibility.

Concrete Examples

This section presents examples of visualizations to illustrate key ideas above. These examples take elements from previous work, such as reviewed in Bull (2020), but we have designed them purely to make key ideas clear.

Figure 2 shows a very simple form of OLM similar to skill-meters. It shows an individual learner's progress for a context where there are seven topics. A learner would see the initial form, in Figure 2a, with a legend and seven gray cells indicating there is currently no data about the learner's progress. Figure 2b shows the OLM after the learner has completed a learning activity (the legend is not included in the screenshot). Now the cells for Topics 1, 2 and 5 have become blue, with Topic 1 indicating mastery and the other two indicating beginner level. Figure 2c now displays the value of cells which have changed since the previous stage. The next view of the OLM, in Figure 3a is after the user has completed the next learning activity.

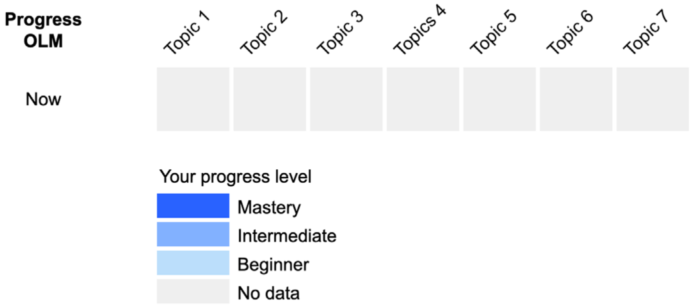


Figure 2a. Initial Visualization

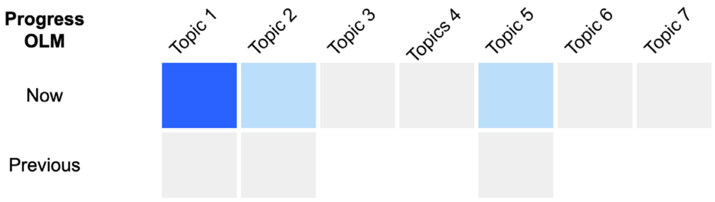


Figure 2b. After learner completed the first activity

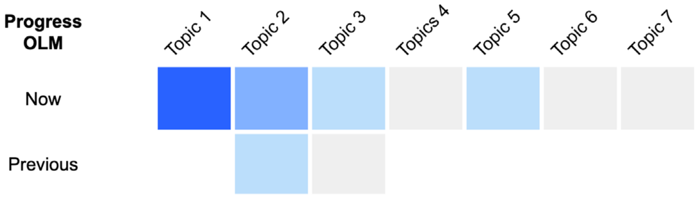


Figure 2c. After learner completed the second activity

Figure 2. OLM visualization to enable a learner to answer: *Am I making progress?* This was designed for quick thinking - at-a-glance. It shows three stages (a) at start up. It also shows the legend which is not included in (b) or (c) which show the changes after learning activities.

The main row of cells, each showing the current state of the learner model is similar to the many skill-meters in AIED systems (Corbett & Anderson, 1994; Guerra-Hollstein et al., 2017; Long & Aleven, 2017; Woolf, 2010) and OLMs for independent learner models (Bull, 2020; Bull & Kay, 2007, 2016).

One key decision about the design of this part of the visualization is the choice of the number of levels to display. In this example, we chose three levels (mastery, intermediate and beginner) in addition to the gray for no data. There are two important reasons for this choice. The first is based on the requirements for user perception of the levels for fast-thinking, at-a-glance interpretation (Kay et al., 2020). For this, it is important to have a small number of levels that the user can easily distinguish.

The second reason relates to a core issue in learner modeling; there is *uncertainty* in modeling learning. There are several factors that contribute to this uncertainty. The data about the learner is incomplete, noisy, and uncertain. This is unavoidable. This is because the designer of a learning interface needs to make design trade-offs between the amount of time a learner spends doing assessment tasks and the need for accuracy in the model. For example, it is well known that people make both slips and errors in learning activities. One way to reduce the impact of slips is for the learner to do repeated assessments. But this may be a waste of previous learning time. Beyond the uncertainty in the raw data, there is uncertainty in how to reason about a collection of evidence to conclude how well the learner knows a particular topic. A substantial body of AIED work has explored many ways to do this. Some effective systems have used very simple methods, such as an average of scores on a small number of recent assessment activities. In light of all this uncertainty of the whole modeling process, it is only meaningful for a visualization to depict a small number of values. We still need research to determine whether learners interpret this as indicative of the uncertainty. There has been valuable work that has invited the learner to self-rate their knowledge and their certainty in that assessment and then to see the system's corresponding assessment (Al-Shanfari et al., 2017).

We now consider the second row of cells in Figures 2b and 2c. They enable a learner to see how much their learner model changed in light of the data from the last learning activity. We have not seen this in OLMs but included it so that the learner can readily see what changed in the learner model. If the OLM is automatically updated when the learner completes an activity (or a part of an activity), an on-screen visualization could have some movement to draw the user's attention to the change. Alternatively, the interface could require the learner to click the OLM to make it update. In either case, the design in Figure 2 enables the learner to easily see whether there has been any change (indicated by the presence of any cells on the *Previous* row). In a classroom setting, this could be valuable for a teacher who could move around the class, glancing at the OLMs to see both the current state and the recent changes. This aspect of the design is aligned with more general design principles (Zapata-Rivera et al., 2020) for ITSs.

We have briefly described how the design of the visualization in Figure 2 was driven by the user goal to answer our first question in Table 1, *Am I making progress?* This design may also support the second question, *Am I meeting my own goals?* This will be the case if the learner can make sense of the three levels and align them with their personal goals.

Figure 3 shows how a version of OLM designed to enable the learner to answer the third question in Table 2, *Am I meeting external goals?* The first row of cells is identical to that in Figure 2b. The second row shows the performance of a particular population the user selected to compare themself against. Any such visualization comes with the well known risks of social comparison (Hanus & Fox, 2015; Khan & Pardo, 2016) but people frequently use it to assess themselves (Festinger, 1954) and it has been used in OLMs (Brusilovsky et al., 2015). For this chapter, it is important as one example of a visualization designed for slow, considered thinking (Kay et al., 2020). The remaining questions in Table 1 also require slow thinking. Additional design elements are needed for each of them.

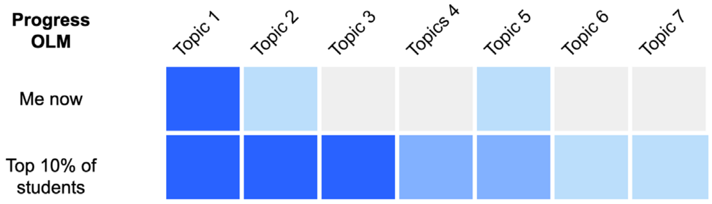


Figure 3. OLM visualization to enable a learner to answer: *Am I meeting external goals?* This was designed for slow thinking as the learner carefully compares their progress with that of students in the top 10% of the cohort.

These examples illustrate the way that the benchmark questions can be used to drive design of visualizations of learning data and OLMs. This reflects a user-centered design approach that starts by determining which of the questions are important for each of the stakeholders. Once that foundation has been established, standard user-experience and Human Computer Interaction (HCI) methods (Hartson & Pyla, 2012) need to design, implement and evaluate the visualizations to assess whether they do enable the stakeholders to actually answer these questions.

Links to GIFT

GIFT has been in active development as a research project since 2010. GIFT has many goals which include being able to be used for research, as well as in a classroom environment. Due to constraints that come from attempting to cover all domains, and all uses of ITSs, there are certain elements of GIFT that are further developed than others. Specifically, the tools that are of use to researchers generally require less interface design and have less demanding usability considerations because the researchers have high digital and data literacy. In contrast, interfaces that are meant for less technical end users have not yet been adequately addressed in many cases. For instance, there is an ability for an instructor to extract data from their student's performance in the ITS. However, it requires using a data extractor tool and selecting the information to include in the output report. Many of these options for inclusion in the report are highly technical, and do not necessarily use terminology that is aligned with that of instructors. There is a current gap, and an opportunity to implement data visualizations in GIFT, to help support the understanding of student performance for both the instructor, as well as the student themselves.

Examining data visualizations that have been created for other ITSs and frameworks is beneficial for the GIFT project. In particular, it is important for commonalities between systems that cover many different topics to be examined to allow GIFT to retain flexibility in the topics that are covered, while also adding utility for instructors and students. It is important for GIFT to make design choices that consider Figure 1, about the desired level view of students that can be examined, and the approach to be used. In the case of GIFT, as it is generalized, both data at the student level, as well as the cohort level is highly applicable. It may also be relevant to use the defined learner model as a structure for presenting information to the instructor, such as how students performed on concepts and topics, errors that they made, and number of attempts to master the material. A further consideration of the design of data visualization tools for GIFT is for team training, and how the learner is performing themselves, as well as their contributions to a team task, and the overall team's performance.

As with the other tools that were created for GIFT, the approach that will likely work best for developing data visualization tools within GIFT is to start with a traditional classroom focus which focuses on individual student and cohort performance. This initial design can be made with future goals for team tutoring in mind, but can produce a fully functional and realized data visualization tool that addresses the performance of individuals and classes.

The SWOT Analysis of Data Visualization in ITSs will be highly relevant to GIFT's development of data visualization tools and interfaces. The highly generalized nature of GIFT is challenging as it requires the system design to remain as flexible as possible. It also requires the GIFT tool development to take into account visualization approaches of many different systems in varying topic areas.

References

Al-Shanfari, L., Demmans Epp, C., & Baber, C. (2017). Evaluating the effect of uncertainty visualisation in open learner models on students' metacognitive skills. *International Conference on Artificial Intelligence in Education*, 15–27.

Bodily, R., Kay, J., Aleven, V., Davis, D., Jivet, I., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: A systematic review. *Proceedings of the Eight International Learning Analytics & Knowledge Conference*, 1–10.

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., & Zadorozhny, V. (2015). The value of social: Comparing open student modeling and open social student modeling. *International Conference on User Modeling, Adaptation, and Personalization*, 44–55.

Bull, S. (2020). There are open learner models about! *IEEE Transactions on Learning Technologies*, *13*(2), 425–448.

Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI: Open learner modelling framework. *International Journal of Artificial Intelligence in Education*, *17*(2), 89–120.

Bull, S., & Kay, J. (2016). SMILI: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *I. J. Artificial Intelligence in Education*, *26*(1), 293–331. https://doi.org/10.1007/s40593-015-0090-8

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.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.

Guerra-Hollstein, J., Barria-Pineda, J., Schunn, C. D., Bull, S., & Brusilovsky, P. (2017). Fine-Grained Open Learner Models: Complexity Versus Support. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 41–49. https://doi.org/10.1145/3079628.3079682

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152–161.

Hartson, R., & Pyla, P. S. (2012). *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.

Kay, J., Bartimote, K., Kitto, K., Kummerfeld, B., Liu, D., & Reimann, P. (2022). Enhancing learning by open learner model (OLM) driven data design. *Computers and Education: Artificial Intelligence*, 100069.

Kay, J., & Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, *50*(6), 2871–2884.

Kay, J., Rus, V., Zapata-Rivera, D., & Durlach, P. (2020). Open learner model visualizations for contexts where learners think fast or slow. In *Design Recommendations for Intelligent Tutoring Systems: Data Visualization* (p. 11). US Army Combat Capabilities Development Command – Soldier Center.

Khan, I., & Pardo, A. (2016). Data2U: Scalable real time student feedback in active learning environments. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 249–253.

Long, Y., & Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an open learner model. *User Modeling and User-Adapted Interaction*, *27*(1), 55–88.

National Academy of Sciences, Engineering, and Medicine (2018). *How People Learn II: Learners, contexts, and cultures.* Washington, D.C.: National Academies Press.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3), 313–350.

Roscoe, R.C., Craig, S.D., & Douglas, I. (2018)(Eds.), End-user considerations in educational technology design (pp. 205-216). Hershey, PA: IGA Global.

Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Zapata-Rivera, D., Graesser, A. C., Kay, J., Hu, X., & Ososky, S. J. (2020). Visualization implications for the validity of intelligent tutoring systems. *Design Recommendations for Intelligent Tutoring Systems*, *Volume 8 – Data Visualization. US Army Combat Capabilities Development Command Soldier Center*, 61.

Chapter 11 – Competency-Based Scenario Design In Intelligent Tutoring Systems SWOT Analysis

Patrick Kyllonen[1], Robby Robson[2], Judy Kay[3], and Bob Pokorny[4]
Educational Testing Service[1]; Eduworks Corporation[2]; University of Sydney[3]; Affinity Associates LLC[4]

Introduction

Overview

This chapter defines competency-based scenario design (CBSD) and discusses strengths, weaknesses, opportunities, and threats (SWOT) to this approach. We define scenarios and competencies from different perspectives, and we discuss implications of these definitions for the concept and scope of CBSD. We then review the SWOT analysis. Strengths include that CBSD supports training approaches that have proven efficacious and CBSD can identify critical training needs. Among the weaknesses is that few training systems are currently designed to take advantage of CBSD so that high quality efficacy evidence for CBSD training is limited. Among the opportunities are that human tutoring works, when implemented well, and the ITS (intelligent tutoring system) promise has always been to meet human tutoring standards. Threats include the danger of overhyping. We suggest how the Generalized Intelligent Framework for Tutoring (GIFT) might improve, by demonstrating value in the market, and by expanding the kinds of assessments that can be easily accommodated, such as situational judgment tests and collaborative problem-solving tasks, as well as the provision of technology for scoring ill-defined, subjective, and complex tasks. We conclude with a discussion of the benefits of fielding a widely used competency training system for general competencies to get feedback on the nature of successful competency performance.

Background

CBSD is an approach to designing instruction that teaches and assesses competencies in scenarios. A competency is “the set of skills, knowledge, abilities, attitudes, behaviors, and habits of practice required in the performance of an activity or task within a specific context" (IEEE, 2022)[2].

A *scenario* is simply a description of a sequence of hypothetical events or “imagined future” events (Mor, 2013, p. 195). But in software design for human-computer interaction (HCI), scenario-based design is a set of techniques in which the system future state is described at an early point in the design process (Rosson & Carroll, 2012). Scenarios are stories that involve an agent and a goal. Scenarios describe the setting, participants (actors and agents), tasks, goals, objectives, and capabilities of the participants, tools available, and actions and events that lead to an outcome (Carroll, 2000). In the training context, scenarios might represent tasks or problems that need solving: “scenario-based learning (SBL) is a method of using specifically designed scenarios for interactive teaching <affording> immediate feedback, team collaboration, real-world thinking, and application in a safe environment where negative outcomes do not harm stakeholders or equipment” (Allen Interactions, 2022). A training scenario might involve training

[2] Similar definitions, from an organizational psychology perspective, have been expressed by Fleishman et al. (1995), Bartram et al. (2002), Bartram (2011), and Campion et al. (2011). Chouhan and Srivastava (2014) (also Stevens, 2012) provide histories of definitions and adopt Bartram’s (2011).

objectives (e.g., competency development), a scenario narrative, and trainee tasks or observations (i.e., assessments), with optional storyboards and tools (Graffeo et al., 2015, p. 1488).

By our definition, the following training systems are ones that have included elements of competency-based scenario design:

> **SHERLOCK** (Lesgold et al., 1988): an early CBSD effort; a practice environment for complex troubleshooting jobs on the avionics systems of F-15 (later, F-18) aircraft in the U.S. Air Force. Trainees are given malfunctioning electronic modules to troubleshoot (see also, Pokorny et al., 2013).
>
> **The Tactical Language and Culture Training Systems (TLCTS)** (Johnson & Valente, 2009): an environment in which students rapidly learn foreign languages and culture through agent interactions, spoken language training, and tutoring. Students carry out a civil affairs mission by entering a town, establishing contact with locals including the local leader, with all of these represented as Artificial Intelligence (AI) characters in an interactive 3d game.
>
> **Negotiation Training** (Johnson et al., 2019). This system trains participants to claim and create value in a negotiation exercise with a virtual partner by exchanging information with their partners, exploring tradeoffs (log rolling), anchoring with early offers, avoiding early concessions, and expressing willingness to walk away, receiving personalized feedback throughout the exercise. A version of this exercise was implemented in GIFT.
>
> **Mission Essential Competency (MEC) training for F-15 Mission Training Centers** (Colegrove & Bennett, 2004): live, simulated, and virtual training for aircrew based on 18 sortie missions. The system is adaptive and competency-based.
>
> **STE Experiential Learning - Readiness (STEEL-R)**, which gathers data from synthetic, semi-synthetic, and live environments and includes a competency-based data strategy based on GIFT, the Competency and Skills System (ADL, 2020), and a competency-based exercise design tool (Goldberg et al., 2021; Hernandez, et. al., 2022).

CBSD is an approach to developing ITSs. It is a method for designing a certain kind of training experience (scenarios, GIFT's domain module) for a certain kind of training (competency training, in GIFT's pedagogical module) to elicit certain kinds of evidence (evidence for competence in GIFT's learner module).

Prior to presenting the SWOT analysis and lessons that can be brought to GIFT development, we first examine in more depth definitions of competencies and scenarios. Our goal is to draw ideas from diverse literatures to help inform GIFT development.

Competencies

The *competencies* movement, favoring mastery evidence over seat time, has been influential in education and training.

Higher Education

In higher-education, the Carnegie Unit and the credit hour are firmly established in the American education system to measure progress toward degrees, but there are calls for reform towards competency-basic metrics (Silva et al., 2015). Demonstrating competencies can be an alternative to accumulating credits as a means for degree attainment. As Western Governor's University (WGU) (2022) puts it in their promotion materials:

> What is competency-based education? Simply put, it measures skills and learning rather than time spent in a classroom. Students progress through courses as soon as they can prove they've mastered the material, rather than advancing only when the semester or term ends.

At WGU students access courseware and demonstrate competency through objective assessments or graded papers or presentations. *Competencies* are the knowledge and skills that would be acquired from taking a course, and courses and competencies are much the same as in traditional higher education, as are the types of degrees issued. Southern New Hampshire University, a prominent competency-based example, issues undergraduate, graduate, and MBA degrees in over 180 traditional academic programs such as history, economics, and nursing. McClarty and Gaertner (2015, p. ii) point to the crucial role assessments play in competency-based education (CBE). They argue that the "viability of CBE programs hinges on the credibility of these programs' credentials in the eyes of employers. That credibility, in turn, depends on the quality of the assessments CBE programs use to decide who earns a credential."

K-12 Education

In U.S. K-12 education, too, CBE emphasizes evidence of mastery over seat time. CBE is defined by meaningful assessment that provides actionable evidence, personalized pathways with learning progressions, varied pacing, emphasis on transferable skills (e.g., problem-solving, creativity, collaboration), differentiated (i.e., personalized) support, equity strategies, and rigorous learning expectations (Levine & Patrick, 2019; Tan et al., 2017). General principles exist for defining competencies—"explicit, measurable, transferable learning objectives" (Freeland, 2014, p. 13), but no specific competency-defining procedure is used and instead the process is decentralized with different schools and districts defining them in different ways (Freeland, 2014, p. 13-15). Still, there is considerable discussion of the benefits of CBE targeted to practitioners (Reich & Huttner-Loan, 2020). There are also proposals for assessing competencies (Burrus et al., 2023). But little research has been conducted thus far on CBE efficacy and implementation (an exception is Steiner et al., 2015, who reviewed personalized learning).

Medical Education

Medical education has embraced the competency model particularly in graduate medical education (i.e., post M.D. or D.O. receipt in the U.S.) with competency-based residency as an alternative to the traditional time-in-training model (Ebert & Fox, 2014; Mills et al., 2020; Stodel et al., 2015). The Accreditation Council for Graduate Medical Education ([ACGME], Al-Temimi et al., 2016) has identified six competencies for which students must provide evidence: patient care, medical knowledge, practice based learning and improvement, systems based practice, professionalism, and interpersonal skills communication, each of which is specified further (e.g., communication includes communicating with patients, with other physicians, as a member of a team, in a consultative role, and by maintaining good records). This illustrates the breadth and nature of competencies—competencies are broad, not narrow, and soft skills are emphasized. Residents provide evidence in the form of Miller's (1990) pyramid (which includes the levels knows, knows how, shows how, does), through multiple-choice tests, surveys, patient surveys, and other indicators, set by the program.

Workforce

In organizations, competencies are used for talent management--aligned recruiting, hiring, training and development, compensation, and promotion and succession management. Because human resources organizations such as Saville & Holdsworth Limited (SHL) (Bartram, 2011; Bartram et al., 2002) have identified a market for assisting other organizations in defining competencies specific to those organizations, it is instructive to review how SHL provides such assistance. SHL created a general competencies framework, drawing from competency models from other organizations (at the time of development, those included Hay, PDI, and Lominger), then tailors the competencies to industry and organization clients.

In defining competencies, SHL (Bartram et al., 2002) highlight some of their specific features.

- Competencies are behavioral repertoires and not the same as knowledge and skills;
    - they also are not the same as competence. Competencies relate to underlying behavior; competence is an attainment level of the competency;
- Competencies can be hierarchically defined. For example, SHL defines the "Great Eight" competencies (Bartram, 2005): leading and deciding, supporting and co-operating, interacting and presenting, analyzing and interpreting, creating and conceptualizing, organizing and executing, adapting and coping, and enterprising and performing. More fine-grained competency breakdowns exist specifying 20 dimensions and further into 112 components.
- Competencies are associated with job types. SHL distinguishes Management, Customer Contact, Directors; However, an even more elaborate job type characterization comes from the U.S.'s occupational database, O*NET (O*NET OnLine, 2022), which differentiates 900 occupations (accounting for all occupations in the U.S.) each of which contains many job titles; O*NET is organized into 19 sectors and hierarchically arranged within sector down to the occupation and job type levels.
- Competencies can be characterized by level of complexity. O*NET (O*NET OnLine, 2022) defines five job zones (Zone 1 to Zone 5), ordered by education and experience requirements.
- The competency framework is then used to build custom, organization-specific competency models.

In education, competencies are defined by the curriculum. In organizations competencies are typically defined through *competency modeling* (or *competency mapping*), a process to determine what workers must be able to do for the organization to be successful (Campion et al., 2011; Chouhan & Srivastava, 2014; Stevens, 2012). Competency modeling identifies competencies linked to business objectives by distinguishing top from average performers; it allows for progressions across levels and is cast in organization-specific language. The process is initiated with articulation of objectives by executive management. This is a contrast to traditional job task analysis, which surveys workers for the tasks they do. Competencies are not lists of knowledge, skills, abilities, and other factors (KSAOs), but start with what the organization wants workers to be able to do, "backing into the tasks and KSAO's" (Campion et al., 2011, p. 227). Campion et al. (2011) define best practices in competency modeling—how they are identified, organized, and communicated, and how they are used for organizational development and alignment of Human Resource (HR) systems.

CareerOneStop (2022), a partner of the Americanjobcenter network and sponsored by the U.S. Department of Labor, has created a competency model clearinghouse, which provides industry models, lists use cases, provides resources, and includes a build-a-model tool that helps organizations build a competency model, based on a competency library. The Society for Human Resources Management ([SHRM], n.d.) has created a Competency Model (defined as a set of competencies) for HR professionals based on 111 focus groups for 1200 HR professionals following best practices guidelines (Campion et al., 2011; Shippmann et al., 2000).[3]

Military

Colegrove and Alliger (2002) defined mission essential competencies as higher order, job-contextualized functions less general than ones found in typical business environments (Tossell et al., 2006). Like the seat time versus mastery issue in CBE and Computer Based Training (CBT), Colegrove and Bennett (2004 abstract) argued that competency-based aircrew training emphasizes "the required proficiency rather than the number of times the mission has been performed." They suggest that competency-based training provides "the ability to compare individual aircrew performance to a defined proficiency level, maintain acceptable levels of performance and target areas requiring improvement," thus "focus(ing) on mission performance rather than mission type" (p. 2). Thus, the Colegrove and Bennett (2004) application to aircrew training shares the perspective from organizational psychology that competencies are defined with respect to organizational objectives ("mission performance") and based on outcomes ("required proficiency"), and that there is a clear pathway to improved proficiency ("target areas requiring improvement").

Psychometrics and Measurement Science

As noted in both the competency-based education (McClarty & Gaertner, 2015) and competency-based training literatures (Campion et al., 2011, p. 229), assessment plays a crucial role in establishing competency proficiency levels—seat time is downplayed and competency evidence is highlighted. Thus, competency-based design will by its nature emphasize assessment due to its reliance on assessed proficiency rather than experience counts and other proxies. As such, it is useful to consider psychometrics and measurement science perspectives on proficiency assessment as there are numerous psychometrics frameworks and concepts that are relevant to the task of assessing proficiency. We suggest several possibly relevant concepts here.

Performance standards (and performance assessments, performance level, performance-level descriptors)

The *Standards for Educational and Psychological Testing* American Educational Association, American Psychological Association, National Council on Measurement in Education [AERA, et al.], 2014) (hereafter, The *Standards*) defines performance standards as "descriptions of levels of knowledge and skill acquisition contained in content standards, as articulated through performance level labels (e.g., "basic," "proficient," "advanced"); statements of what test takers at different performance levels know and can do; and cut scores or ranges of scores on the scale of an assessment that differentiate levels of performance." (p. 221). The *Standards* provides similar definitions for related concepts. There is a large literature on the use of performance standards and related concepts for designing assessments, setting standards and cutpoints, and reporting (Czisek, 2006; Hambleton & Pitoniak, 2006; Lane & Stone, 2006).

[3] SHRM identified 9 competencies: Business Acumen, Communication, Consultations, Global and Cultural Effectiveness, HR Expertise, Leadership & Navigation, Relationship Management, and Ethical Practice, that can be used to identify strengths and areas for growth, design professional development activities, design talent acquisition plans including selection assessments, identify department strengths and gaps, and communicate the role HR can play within the organization.

Formative versus reflective latent variable models and network psychometrics.

Reflective latent variable models state that a latent variable, such as cognitive ability or a personality trait causes item responses (and test scores). Such models are the foundation for concepts such as reliability and factor analysis, classical test theory and item response theory. However, competencies may be multidimensional constructs (e.g., the social and intellectual parts of leadership may be independent; *can do* and *will do* aspects of competencies may be independent) and so reflective, unidimensional latent variable models are not suited to determining psychometric properties of competencies, such as reliability or validity (Edwards & Bagozzi, 2000). Alternative specifications such as multidimensional item response theory (MIRT) (Reckase, 2009), formative latent variable models (Bollen & Bauldry, 2011) or network psychometrics (Borsboom, 2022) may be useful. Wang et al. (2018) demonstrate how cognitive diagnostic modeling (CDM) can be used in student modeling, and Deonovic et al. (2018) show correspondences between item response theory and Bayesian Knowledge Tracing.

The reason dimensionality is important from a training perspective is that training on one aspect (or dimension) of a multidimensional construct would not be expected to transfer to a different aspect (or dimension)—training the social aspects of leadership would not transfer to the intellectual decision-making aspects. But training on any aspect of a unidimensional construct would be expected to transfer to other correlated aspects of that construct—training on a word processing system transfers to other word processing systems (Singley & Anderson, 1989). This is not to say that there is not transfer across positions; there is (Gathmann & Schönberg, 2010). But that is because to be successful at a position requires learning all aspects of that position. Initial training per se on one aspect will not transfer to an independent aspect.

Learning progressions/Learning trajectories

Learning progressions (LPs)[4] "are theories that describe students' knowledge and skills of a certain content area in an increasing order from simpler to more sophisticated" (Pham, 2019, p. 28); or "hypothesized descriptions of the successively more sophisticated ways student thinking about how an important domain of knowledge or practice develops as children learn about and investigate that domain over an appropriate span of time" (Corcoran et al., 2009, p. 37). Also, "most students' understanding will move through these intermediate conceptions in roughly the same order, though perhaps at quite different rates…" (Corcoran et al., 2009, p. 42). Learning progressions are different from stage theories, such as Piaget's (1952) and Kohlberg's (1958) in that they can be seen as "modal paths, meaning paths that most students take" (R. Bennett, personal communication, November 4, 2022) rather than necessary developmental stages (Confrey, 2018, p. 9). Learning progressions or learning trajectories are useful because, theoretically, they can serve as testable models of learning (Choi & Mislevy, 2022). Practically, they can provide the basis for curriculum materials and practices, instructor training (professional development), and diagnostic assessments (Confrey, 2018, p. 14). The psychometrics framework of CDM is suited to accommodate learning progressions (Chen et al., 2017; de la Torre & Douglas, 2004; Kizil, 2015; Rupp et al., 2010); also, hidden Markov models and dynamic Bayesian networks (Choi & Mislevy, 2022; von Davier et al., 2021) are suited to test as well as possibly to discover learning progressions (Jia et al., 2021).

Hierarchical relationships

In both Bartram (2011) and Campion et al.'s (2011) characterizations, competencies exist at multiple levels of abstraction from a low level (Bartram proposes a competency system with 112 components) to a high

[4] Confrey (2018) points out that *learning progressions* as a term dominates in science education, but that in mathematics education *learning trajectories* is more common. The latter term may denote more finer cognitive distinctions, but the two terms are often treated interchangeably.

level (Bartram [2011] proposes 8; Campion et al. [2011] suggest "a few"). A psychometric measurement system must be capable of modeling and reporting on competencies at multiple levels in the hierarchy. Kay and Lum (2004) propose ontologies (essentially, hierarchical student models) to enable *ontological inference*, inferring student knowledge at one level based on evidence from a different level. Shute and Zapata-Rivera (2012) present a related proposal. Within psychometrics, a similar function is served by hierarchical item response theories (Johnson et al., 2006; Rijmen, 2011), hierarchical cognitive diagnosis models (de la Torre & Douglas, 2004; Ma et al., 2022), and network approaches that can handle hierarchies (Borsboom, 2017; 2022). Perhaps the fundamental difference between proposals such as Kay and Lum (2004) and psychometrics approaches are that the latter approaches model measurement error using identified models and evaluates model fit to the observed response data. This has implications for the confidence one has in assertions about student knowledge based on response evidence.

Scenarios

There are two senses of *scenario* in CBSD. Scenarios, or design scenarios, are an approach to human-computer interaction (HCI) design, an alternative, at least in emphasis, to the functional design approach ("rational decomposition into features and functions," Carroll, 2000, p. 316) traditionally used in software development. Scenarios can be stories, simulations, use cases, or storyboards (Alexander & Maiden, 2004), and their role in system development is to highlight how users will interact with the system, what their needs will be; and to enable early prototyping and system evaluation (Carroll, 1995): "…they are simultaneously concrete and incomplete; they are at once dreams, design arguments, scientific analyses, and software specifications…they are all about us by being all about the contexts within which we experience and act" (Carroll, 2000, p. 316).

Training applications of scenarios may refer to this software development aspect, but they also refer to two other roles of scenarios: as providing learning and practice opportunities, and as contexts eliciting evidence for the skills or competencies being taught. Consider the following definitions of scenarios, simulations, and scenario-based learning:

- SCENARIOS: "information-rich task/problem contexts that are closely aligned with real-world situations that professionals face on their jobs….rather than stripped-down abstract simplifications." (Sinatra et al., 2022, p. vi).

- SIMULATIONS: "an educational tool or device with which the learner physically interacts to mimic an aspect of clinical care for the purpose of teaching or assessment" (Cook et al., 2013, for health professions education).

- SIMULATIONS (second definition): "approximations of practice in which the complexity is reduced … can help engage learners in specific aspects of professional practice and are promising in order to avoid confusion and efficiently use resources for learning and instruction. These approximations of practice can be realized in higher education with simulations, which allow students to use authentic problems and also to create a learning environment to practice and facilitate the acquisition of target complex skills…" (Chernikova et al., 2020, higher education).

- SCENARIO-BASED LEARNING: "SBL (also called "problem-based learning" or "case-based learning") is an instructional environment in which participants solve carefully constructed, authentic job tasks or problems. While solving the problems, they are carefully guided to learn the associated concepts, procedures, and heuristics of expert performers." (Clark, 2009).

Simulations, scenarios, and scenario-based learning are not identical concepts, but they overlap considerably and it is useful to consider them together. Having realistic scenarios is generally useful in training but Sinatra et al. (2022) also point out that a realistic scenario might sometimes not be useful because the real-world situation has changed or the scenarios are unable to elicit evidence for a competency being trained.

Scenarios in the Military

There is a long history in military settings of using scenarios to simulate performance environments for training, including combat and air training environments. Tossell et al (2006) describe an application of the United States Air Force's Mission Essential Competency (MEC) framework to training in the Air and Space Operations Center (AOC). Training is conducted at many levels, such as classroom training, large-scale exercises (e.g., Blue Flag exercise), on-the-job training, and most recently, in the synthetic training environment (STE) being developed by the U.S. Army Futures Command. There, AI generated simulations "not only construct and replicate tough and realistic scenarios for Soldiers, but also collect detailed data on how Soldiers react under pressure, further informing training needs and operational planning methods and continually increasing training thresholds" (Thompson, 2022). While projects such as STEEL-R (Goldberg et al., 2021; Hernandez et. al., 2022) look to integrate CBSD into synthetic and live environments in the future, CBSD is currently already in extensive use in part-task trainers or simulators, which provide focused training on key competencies, particularly ones that have been identified as essential and critical to improved trainee performance. In this scheme, simulated training scenarios or vignettes are formed from collections of tasks, which in turn derive from training objectives and mission essential competencies, knowledge, and skills. Specifically, what are referred to as *developmental experiences* are shorter, focused scenarios designed to provide the greatest number of knowledge and skill elements in a simulation, that is, rich scenario events providing targeted opportunities for learning.

Psychometrics and Measurement Science

Psychometrics, particularly item-response theory (IRT), assumes that responses are a function of both person and item characteristics, providing a framework for modeling responses in scenarios. Scenarios reflect complex items, with multiple potential influences on item responses, such as knowledge of the context as well as a requirement for the skills that the construct or sets of constructs being evaluated represent. Because of the multidimensional nature of scenarios, cognitive diagnostic models (CDMs) (Xin et al., 2022) may be particularly well suited for the modeling of scenario responses. With CDMs, one specifies the constructs that may be invoked in a set of scenarios (or items) in the form of a Q-matrix, a binary matrix of items by item-required attributes. These binary item-attribute relations are typically specified in advance by experts (Culpepper, 2019) in a confirmatory sense (i.e., 1 if the attribute is required on the item, 0 otherwise), but exploratory approaches that discover appropriate Q-matrices are also possible (Ma & Hu, 2021). CDMs also include capabilities for modeling growth and change in competencies over time (Kaya & Leite, 2017).

Some other key concepts relating to the psychometrics of scenario response modeling are reliability and construct irrelevant variance. Reliability is the means "to quantify the precision of test scores and other measures" (Haertel, 2006, p. 65). One way to increase reliability is to increase test length, or as applied here, the number of scenarios in which a trainee would participate (assuming each scenario elicits a response) or the number of response-eliciting events within a scenario. With a test like a vocabulary test, it is straightforward to increase reliability, r, by increasing by a factor of k the number of vocabulary items (for example, $k = 2$ increases reliability to $2r/[1 + 2r]$). It may not be as simple to double the number of scenarios or response-eliciting events, depending on the nature of the scenario.

A SWOT Analysis of Competency-Based Scenario Designed Instruction

This background provides a definition of competency-based scenario design (CBSD) from the perspective of different research and practice communities. CBSD is several things: it is a scenario approach to designing HCI environments, a training approach that uses scenarios to train on, and a way to design competency-based education or training. We identify strengths, weaknesses, opportunities, and threats related to CBSD instruction from these different perspectives.

Strengths

CBSD supports training approaches that have proven efficacious. CBSD, as discussed in this chapter, supports ITS training that uses scenarios. But do ITSs and scenarios produce learning gains relative to a typical-instruction baseline. The definition of ITSs varies (Kulick & Fletcher, 2016)—a traditional definition reflects GIFT structure (an ITS is one that includes domain, pedagogical, and student models); an alternative defines ITSs as providing prompting, hinting, and support feedback during rather than only after problem solving (VanLehn, 2011). Defining ITSs with this latter definition, combined with author and expert ITS designation, Fletcher and Kulik (2015) found that ITSs were efficacious, relative to conventional instruction, although with considerable variability. They estimate an average effect size of .66 *SD*, but as Nickow et al. (2020) point out, Kulik and Fletcher included non-experimental studies and lab studies, which likely show larger effect sizes than field studies. Ma et al. (2014) and Nesbit et al. (2014) similarly found evidence for ITS efficacy (.43 *SD* for Ma et al., 2014) for ITSs. These studies did not include (or identify) scenario ITSs, but scenario training, when defined as simulations, also has been shown to produce gains. Simulation training (compared to non-simulation instruction) has been shown to improve outcomes in medicine (Lorello et al., 2014) and in higher education more generally, moderated by including reflection (more beneficial for high prior knowledge learners) or scaffolding with examples (relatively more beneficial for low prior knowledge learners) (Chernikova et al., 2020).

CBSD can identify critical training needs. CBSD leads to systems that train competencies, and many aspects of the competency modeling process are useful for identifying the most important training requirements for an organization. Competency modeling starts at the top, considers the organizational context, links competencies to organizational goals and objectives, including future objectives, and promotes discussions about proficiency levels. The process can be rigorous in identifying competencies and levels, using traditional job analysis techniques such as subject matter expert (SME) interviews, brainstorming with focus groups, survey methods, critical incident techniques, and employee surveys that get ratings of competency importance and the degree to which the competency differentiates high from average organizational performers (Campion et al., 2011). This process for identifying competencies ensures that training focuses on professional development that will be the most important to an organization.

Competencies are the right grain size for training. The competency modeling process, particularly as implemented in workforce settings, ensures that important competencies are the ones selected for training. The process also is designed to identify the right grain size for training, and for Human Resources (HR) systems alignment generally (for personnel selection, compensation, promotion, as well as training). Because through competency modeling managers have articulated the organizational need, developing training to serve that need is relatively easy to justify, as is competency-based evaluation of the training and trainees.

If the goal is to train on scenarios, which CBSD is for, scenario design is a useful approach. This pertains to scenario design as an HCI strategy (Carroll, 2000). Scenario settings, actors, goals, intentions, actions, and resolutions can be discussed with designers and users (instructors and trainees) early in the design process. For example, Alliger et al. (2004) describe the process of developing scenario training through a series of workshops involving SMEs focused on the competencies for a position, and the specific training needs for that position. Such discussions can involve preparing prototypes using sketches, storyboards, wireframes, videos, and rapid prototyping tools. Scenarios force attention on contextual details and temporal dynamics. Scenarios are flexible representations tied closely to use. They are easily developed, shared, manipulated, and categorized for later use. They can be used for ideation in scenario-mapping exercises.

Scenarios allow trainees to practice on rare but important or critical events. Lesgold (2012) points out that one of the benefits of scenarios and simulations is that they allow trainees to practice on rare but important or critical events. These can be safety related, as in emergency procedures or working in hazardous environments, or just opportunities to practice competency-related skills that are not typically required, such as diagnosing or troubleshooting rare conditions. Lesgold (2012) argues that this feature of scenarios alone can often justify the expense of training system development. Scenarios provide an environment to prepare people to respond effectively to events one hopes do not occur but that require rapid action in high-pressure situations.

Scenario training promotes transfer of training. By their nature, scenarios are designed to reflect aspects of the real-world environment for which the competency is being trained. The similarity between the practice environment and the real-world environment should affect transfer of training. Transfer of training—transfer is a function of similarity—is a powerful, reliable, and general phenomenon operating through skill transfer settings ranging from narrow cognitive skill acquisition (Singley & Anderson, 1989) through transfer from workplace training to on-the-job performance (Blume et al., 2010; Ford et al., 2018) to shifts in career paths (Gathmann & Schönberg, 2010; Robinson, 2018).

Weaknesses

CBSD supports training approaches that sometimes produce results no better than typical instruction. Several meta-analyses have shown small or no benefits for ITSs. Steenbergen-Hu and Cooper (2013) found negligible ITS effects on mathematical learning, although Kulick and Fletcher (2016) attribute that finding at least partly to Steenbergen-Hu and Cooper's (2013) overly broad definition of an ITS, which included less effective instructional systems. A review of studies of Carnegie Learning's Cognitive Tutor curriculum concluded "mixed effects," "no discernible effects," and "potentially negative effects" on Algebra, general mathematics achievement, and geometry, respectively, in addition to concluding that several studies conducted provided "no evidence" due to their design (What Works Clearinghouse [WWC], 2016, Table 1). A review of competency-based education (CBE) found that little research has been conducted thus far on CBE efficacy and implementation (Steiner et al., 2015, reviewed personalized learning).

There have been few studies producing high quality evidence for the efficacy of CBSD training, or ITS training for that matter. An issue here is that the U.S. Department of Education has stringent standards of evidence[5]. Their three tiers of positive evidence are Tier 3 Promising (requires positive and no negative effects from well-designed experimental or quasi-experimental research), Tier 2 Moderate (requires additionally multiple sites and large samples [$N > 350$], and most meta-analytic weight being based on findings rated "Meets WWC Standards Without Reservations"), and Tier 1 Strong (additionally requires experimental research, that is, randomized control trials). Thus, positive conclusions about educational interventions, such as those cited in the *Strengths* section of this chapter, may at least partially be attributable to relaxed evidence standards; many studies included in the meta-analyses do not meet WWC

[5] The standards are also evolving. The cited document was reviewed under WWC Procedures and Standards Handbook Version 2.0 (2008), whereas the Current Standards Version is 5.0 (August, 2022). See https://ies.ed.gov/ncee/wwc/handbooks

standards. Training research is equally susceptible to relaxed evidence standards and has the additional burden of not being enforced with anything like a WWC evaluation.

Defining competencies is challenging and there are no standards. Competency-based education is not new—Gallagher (2014, p. 18) traced it back to the 19th century and argues that there always has been "an individualized approach to education in which students demonstrate the acquisition of predetermined competencies, typically in a self-paced manner and through performance assessments." He also pointed out weaknesses in the concept of competency-based education, such as the difficulties in defining competencies and levels, the lack of evidence for competency-based education efficacy, the lack of support from teachers and students, the simultaneous proliferation and narrowing of competencies over time, and the finding that students most in need are the least likely to benefit--disadvantaged students are often the first to drop out. In the workforce context, Stevens (2012) similarly points out that competencies are difficult to define and there is a lack of rigor in developing competency models and in assessing competencies.

CBSD training is still relatively novel. Although there are by now several meta-analyses on ITS training, cited in this chapter, and a few on simulation-based training, also cited, we could find no meta-analyses of CBSD training, indicating relative novelty. Where strong efficacy evidence exists, adoption is facilitated, but with no evidence, adoption will be more difficult. The introduction to new ways of doing things is always challenging, and there is a burgeoning science and best practices surrounding managing organizational change, dealing with change management models and overcoming obstacles, such as employee resistance, communications, turnover, and costs (SHRM, 2017). As Lesgold (2012) points out, the decision to adopt or develop a new training system is risky for a training director comfortable with familiar approaches, and less comfortable with new technology, particularly when it requires a significant financial investment.

It is important to avoid "Teaching to the Scenario." "Teaching to the test" or item-teaching (as opposed to curricular-teaching) is decried in education for raising test scores but not the underlying knowledge and skills (Popham, 2001). It can lead to a narrowing of the curriculum (Levin, 2012) and depressed long-term performance and interest in the subject (Carrell & West, 2010). It is possible that teaching to the scenario could produce similar negative consequences. Sinatra et al. (2022) mentioned the consequences of scenario training when the real-life situation changes, an instance of too-narrow training. If not cost prohibitive, having multiple, diverse scenarios is a mitigation strategy. A related problem is that non-essential scenario features, or, using testing language, *construct-irrelevant* features might introduce learning and responding requirements unrelated to the training construct of interest. A "motivate and inspire subordinates" competency might involve interacting with members whose cultural background the target is unfamiliar with or uncomfortable with. Other examples include the appearance (e.g., race/ethnicity, gender, stigmatized appearance) of a partner in a negotiation task, off-putting scenario aesthetics, or sluggishness in system responsiveness, any of which could affect both training efficacy and inferences drawn from the target's performance.

Opportunities

Human tutoring works, when implemented well, and the ITS promise has always been to meet human tutoring standards. In the most comprehensive study to date of the effects of human tutoring (i.e., one-on-one or small group supplements to classroom instruction) based on randomized controlled trials, Nickow et al. (2020) estimate a pooled effect size of .37 *SD* on learning outcomes, with stronger effects for earlier grades, in school rather than after school, and when administered by teachers rather than by parents. Although more modest than Bloom's (1984) two-sigma claim, Nickow et al. (2020) suggest that implementation and dosage issues may account for much of the variation between studies. The key finding is that tutoring per se is a robust and powerful educational strategy operating across program and study characteristics, and so the ITS aspiration of mimicking human tutoring remains a promising opportunity.

Competency-based scenario training is a relatively new phenomenon, providing first-mover advantage opportunities. Our review reveals an extensive database on ITSs and their efficacy and a considerable database on scenario or simulation training and their efficacy, but much less, if anything, on scenario-based ITSs. No meta-analyses have yet been conducted on this class of training. There are well-developed methods for identifying competencies as discussed, particularly in the workforce, and yet relatively little on targeting scenario training to those competencies. This suggests a first-mover opportunity for CBSD. First-mover advantages are the competitive advantages participants gain by being first to market, including ownership of the intellectual property and the ability to gain a loyal customer base, one reluctant to switch technologies later.

Targeting training to competencies is a natural fit for organizations with competency-aligned HR systems. As discussed in the introduction, competency modeling and competency-based approaches are used to align HR systems so that personnel selection, compensation, promotion, and succession planning are based on the same set of competencies. In this context, aligning training to those same competencies is a natural step. Competencies can reflect a broader range of knowledge, skill, ability and other factors than are sometimes the target of training. Soft skills are amply reflected in competency definitions (Bartram, 2002). Training soft skills is relatively novel, although there are some analyses of the efficacy of such approaches (Arthur et al., 2003; Martin-Raugh et al., 2020), albeit not necessarily conducted within a competency framework. This presents an opportunity.

There is an opportunity to increase the use of psychometric and measurement models in student modeling and evaluation. This is an opportunity for ITSs generally but may be particularly relevant to CBSD. There have been some efforts to date, such as Deonovic et al. (2018) who connects Bayesian knowledge tracing to item response theory models. Another natural connection is with cognitive diagnostic modeling and multidimensional item response theory, which are beginning to be deployed for this purpose (Su et al., 2021; Su et al., 2022).

There are opportunities to increase our understanding of why ITSs and CBSD training works. Although as we have documented, there is evidence for ITS and simulation training efficacy, there are also circumstances producing no discernible instructional effect. We understand some reasons—transfer to content that differs from the precise content tutored is difficult. But generally, we do not have a complete understanding of when, under what circumstances, and why ITS and CBSD training is effective, or most effective. VanLehn (2011) presented hypotheses for why human tutoring is effective: diagnostic assessment, individualized task selection, sophisticated tutoring strategies, dialogues, domain knowledge, motivation, feedback, scaffolding, interaction/construction—but suggested only the latter three, and possibly motivation, were not refuted by evidence. This led to the useful interaction granularity hypothesis, that tutoring is effective to the degree to which it provides feedback within or after a problem-solving step vs. after an answer is entered, with the finding that step-based tutoring was most effective[6]. But there remain puzzling and inconsistent findings – why does control group instruction that relies on "materials derived from ITS interactions" (Kulik & Fletcher, 2016, p. 68), which is not interactive, apparently provide an improvement over traditional control group instruction? Why does implementation matter so much and how can it be fixed? This calls for more research. Development of CBSDs and ITSs and evaluating them in large-scale studies should pay dividends in contributions to our understanding of how individuals and teams learn, barriers to learning and to good implementation, and best training practices. Benefits may accrue to ITS, instructional science, and learning sciences more generally.

ITSs are not yet prominent in corporate training, but the market is huge and growing. As seen in our review, most ITS and simulation applications are found in education, particularly K-12 education and medical

[6] This was counter to an expectation that sub-step tutoring would be most effective, but Kulik and Fletcher (2016) suggested that this might have been due to different kinds of control groups used in the two kinds of studies.

training, along with military training. Corporate training in the U.S. is a $46B market and expected to grow with e-learning training modules. This represents a major opportunity for CBSD.

Threats

There is always the danger of overhyping. The Gartner Hype Cycle refers to a normal course of evolution for innovations initiating with an *innovation trigger* (a new technology gets attention) through to a *peak of inflated expectations* (hype outweighs evidence) then through a *trough of disillusionment* (early adopters complain), a slope of enlightenment (early adopters see benefits) and finally the plateau of productivity (the technology goes mainstream). This is not a scientific model but is supported by enough anecdotal evidence that many tech sector observers expect this kind of transition to mainstream use. Overhyping is not a threat to CBSD or to ITSs specifically but to AI technologies generally (Ciocca et al., 2021).

Comfort with the familiar. Within the education and training industry traditional commercial systems based on traditional training models can be viewed as competitors to an intelligent system approach. Installing new systems requires a financial investment and possibly changes in ways of doing things. Lesgold (2012) points out that resistance is reduced with good documentation, integration with operations, availablity of system training, availability of staff familiar with the new system, and tools to support rapid courseware prototyping and development.

There may be a perception of expensiveness for CBSD systems. Favorable cost-benefit arguments for CBSD use are necessary, which can include the elements of reduced training time, better alignment between the competencies and how they are trained, and the opportunity to enable practice for rare but important cases.

Suggestions for GIFT

A major challenge for GIFT is demonstrating market viability, for the core markets of the K-12, higher education, workforce, and military sectors. Each sector has unique issues. In K-12, executing well-designed, sufficiently statistically powerful studies on the appropriate target populations, in the appropriate target settings, particularly using randomized control trials, is necessary to obtain *moderate* and *strong* efficacy evidence. Establishing such evidence is a requirement for funding from various U.S. Department of Education programs. The other sectors do not post nor enforce comparable efficacy and implementation standards[7] but research consumers in all sectors are becoming increasingly sensitive to issues of the strength of evidence regarding appropriate inferences that can be drawn from studies, including meta-analyses. U. S. Department of Education (2016) guidelines, such as on relevant, evidence-based intervention selection, the use of logic models (theories of action), and attention to implementation issues such as local capacity, are useful for instructional interventions across all sectors, and the trend towards holding research studies to high evidence and implementation standards is likely to continue. GIFT viability will be enhanced to the extent that it is a component of research studies that produce strong efficacy evidence and a well-developed implementation plan.

GIFT could expand the kinds of assessments that can be easily accommodated, such as situational judgment tests and collaborative problem-solving tasks. Stealth assessments can be conducted during the learning activity itself, but there is typically interest in generalization to tasks or environments outside the ones that are the direct target of instruction. In some cases, survey instruments are used to assess learning from CBSD environments despite their limitations. But there is a potential missed opportunity when a rich scenario-

[7] In the workforce, Kirkpatrick and Kirpatrick's (2006) four levels of evaluation—reaction (affect), learning (knowledge), behavior (transfer), and results (return on investment)—are helpful and well known, but not enforced by any regulatory agency.

based design for training is not accompanied by an equally rich assessment environment. To accomplish this, GIFT could provide technology for scoring ill-defined, subjective, and complex tasks. An example would be a system to video record a trainee's performance (e.g., interview, role play), retrievable for scoring using playback (e.g., pause, fast-forward) and annotation and input tools. More generally, support for a wider variety of assessment tools would be helpful for assessing knowledge and skill gains resulting from instruction. This may also require database accommodation for the kind of process data (keystroke, conversation) such assessments may generate (Hao et al., 2015), or developing and validating a catalog of mappings from performance data gathered by GIFT to assertions of skills and competencies, as is done by the STEEL-R architecture.

General Discussion

There is strong interest in competency-based scenario design within the realm of training approaches. The Air Force Research Laboratory recently released a $67M broad agency announcement for innovative research related to "competency definition and requirements analysis, training and rehearsal strategies, and models and environments that support learning and proficiency achievement and sustainment during non-practice or under novel contexts" (AFRL, 2020; pp. 2-3), a clear indication of a market interest.

There are challenges to widespread adoption of CBSDs. Like any new technology, there are potential implementation barriers due to a financial investment requirement and a change to business as usual. The value of such change must be justified with a cost-benefit argument. Here we suggested several components, particularly, increased learning based on strong efficacy evidence and the provision of opportunities to practice on important but infrequent events. An additional challenge is change management, old habits die hard, but with clear and convincing cost-benefit demonstrations these can be overcome.

There are also challenges specific to GIFT, notably, how CBSD can be facilitated through the structure of GIFT. Suggestions here, to position GIFT to serve as a universal portal or to provide extended assessment capabilities, may vary in how difficult they are to achieve, but regardless of what is done, it is important to develop a strategy that will demonstrate clear market advantages to the GIFT framework.

References

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

AFRL (2020). Research methods and technologies for blended live and synthetic personalized learning, modeling and assessment Open BAA. Broad Agency Announcement Number FA8650-20-S-6099.

Al-Temimi, M., Kidon, M., & Johna, S. (2016). Accreditation Council for Graduate Medical Education Core Competencies at a Community Teaching Hospital: Is There a Gap in Awareness? *The Permanente Journal, 20(4*), 16–067. https://doi.org/10.7812/TPP/16-067

Allen Interactions (2022). Scenario-based learning. Retrieved 12-26-2022 from https://www.alleninteractions.com/services/scenario-based-learning.

Alliger, G., McCall, J. M., Garrity, M. J., See, K., Tossell, C. (2004). *Advanced training for commanders: A competency-based approach to training requirements definition for the JFACC.* 2004 Paper No. 1834. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2004).

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: American Educational Research Association.

Alexander, I. F., & Maiden, N. (Eds.) (2004). *Scenarios, Stories, Use Cases: Through the Systems Development Life-Cycle*. Wiley. ISBN: 978-0-470-86194-3

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(2), 234–245. https://doi.org/10.1037/0021-9010.88.2.234

Bartram, D. (2002). The SHL Corporate Leadership Model. SHL White Paper. Thames Ditton: SHL Group plc

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90(6)*, 1185-1203. doi: 10.1037/0021-9010.90.6.1185.

Bartram, D. (2011). The SHL Universal Competency Framework. White Paper. SHL: The CEB Talent Measurement Solution.

Bartram, D., Roberton, I. & Callinan, M. (2002). Introduction: A framework for examining organisational effectiveness. In Robertson, I.T., Callinan, M. & Bartram, D. (Eds.) *Organisational effectiveness: The role of Psychology* (pp. 1-12). Wiley.

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13,* 4-16. http://dx.doi.org/10.3102/0013189X013006004

Blume, Brian D.; Ford, J. Kevin; Baldwin, Timothy T.; Huang, Jason L. (2010). Transfer of Training: A Meta-Analytic Review. Journal of Management. 36 (4): 1065–1105. doi:10.1177/0149206309352880

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. Psychological Methods, 16(3), 265–284. https://doi.org/10.1037/a0024448

Bolsinova, M., Matthieu J. S. Brinkhuis, M. J. S., Abe D. Hofman, A. D., & Maris, G. (2022). Tracking a multitude of abilities as they develop. *British Journal of Mathematical and Statistical Psychology, 75 (3),* 753-778. 10.1111/bmsp.12276.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D. (2022). Possible futures for network psychometrics. *Psychometrika, 87*(1), 253–265. https://doi.org/10.1007/s11336-022-09851-z

Burrus, J. Rikoon, S. H., & Brenneman, M. W. (2023). *Assessing competencies for social and emotional learning.* New York: Routledge.

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64*(1), 225–262. https://doi.org/10.1111/j.1744-6570.2010.01207.x

Careeronestop (2022). Competency Model Clearinghouse. Retrieved 8-26-2022 from https://www.careeronestop.org/CompetencyModel/

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118(3),* 409-432.

Carroll, J. M. (Ed.) (1995). *Scenario-based design: envisioning work and technology in system development*. New York: John Wiley & Sons. ISBN:978-0-471-07659-9

Carroll, J. M. (2000). Making Use: Scenario-Based Design of Human-Computer Interactions. The MIT Press.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika, 82,* 660-692.

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90(4),* 499–541. https://doi.org/10.3102/0034654320933544

Choi, Y., & Mislevy, R. J. (2022). Evidence centered design framework and dynamic Bayesian network for modeling learning progression in online assessment system. *Frontiers in Psychology, 13.* https://doi.org/10.3389/fpsyg.2022.742956

Chouhan, V. S., & Srivastava, S. (2014). Understanding competencies and competency modeling—A literature survey. *Journal of Business Management, 16(1),* 14-22. e-ISSN: 2278-487X, p-ISSN: 2319-7668.

Ciocca, J., Horowitz, M. C., & Kahn, L. (2021, April 6). The perils of overhyping artificial intelligence. *Foreign Affairs, 100(2).* https://www.foreignaffairs.com/articles/united-states/2021-04-06/perils-overhyping-artificial-intelligence

Clark, R., (2009). *Accelerating expertise with scenario-based learning*. Learning Blueprint. Merrifield, VA: American. Society for Teaching and Development

Colegrove, C. M., & Alliger, G. M. (2002). *Mission Essential Competencies^{SM}: Defining Combat Mission Readiness in a Novel Way.* Paper presented at the NATO Research & Technology Organization, Studies, Analysis, and Simulation Panel, Conference on Mission Training via Distributed Simulation (SAS 38), Brussels, Belgium.

Colegrove, C., & Bennett, W. Jr. (2004). *Competency-based training: Adapting to warfighter needs.* AFRL-HE-AZ-TR-2006-0014. Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division. Available at http://www.dtic.mil.

Confrey, J. (2018). Future of Education and Skills 2030: *Curriculum Analysis. A synthesis of research on trajectories/projections in mathematics.* EDU/EDPC(2018)44/ANN3. Paris: Office of Economic Cooperation and Development.

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J., & Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher, 35(1),* e867-e898, DOI: 10.3109/0142159X.2012.714886

Corcoran, T. B.; Mosher, F. A., & Rogat, A. (2009). *Learning Progressions in Science: An Evidence-Based Approach to Reform.* CPRE Research Reports. Retrieved from https://repository.upenn.edu/cpre_researchreports/53

Culpepper, S. A. (2019). Estimating the Cognitive Diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. Psychometrika, 84(2):333–357.

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., & Maris, G. (2018). Learning meets assessment. *Behaviormetrika, 45(2),* 457-474

de la Torre, J., & Douglas, J.A. (2004). Higher order latent trait models for cognitive diagnosis. Psychometrika, 69, 333-353.

Ebert T. J., & Fox, C. A. (2014). Competency-based education in anesthesiology: History and challenges. *Anesthesiology, 120(1),* 24-31. doi: 10.1097/ALN.0000000000000039. PMID: 24158052.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5(2),* 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Fleishman, E. A., Wetrogan, L. I., Uhlman, C. E., & Marshall-Mies, J. C. (1995). In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), Development of prototype occupational information network content model (Vol. 1, pp. 10.1-10.39). Salt Lake City, UT: Utah Department of Employment Security (Contract Number 94-542).

Ford, J. K., Baldwin, T. T., Prasad, J. (2018). Transfer of Training: The Known and the Unknown. *Annual Review of Organizational Psychology and Organizational Behavior, 5(1),* 201-225.

Freeland, J. (2014). *From policy to practice: How competency-based education is evolving in New Hampshire. Clayton Christensen Institute for Disruptive Innovation.* Retrieved on 11/16/2022 from https://www.christenseninstitute.org/wp-content/uploads/2014/05/From-policy-to-practice.pdf

Gallagher, C. W. (2014) Disrupting the Game-Changer: Remembering the History of Competency-Based Education. *Change: The Magazine of Higher Learning, 46:6,* 16-23, DOI: 10.1080/00091383.2014.969177

Gathmann, C., & U. Schönberg, U. (2010). How general is human capital? A task-based approach. *Journal of Labor Economics 28 (1),* 1–49.

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 *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*, Orlando, FL. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2021&AbID=97151&CID=862

Graffeo, C., Benoit, St. T., Wray, R. E., Folsom-Kovarik, J. T. (2015). Creating a scenario design workflow for dynamically tailored training in socio-cultural perception. *Procedia Manufacturing, 3,* 1486-1493. Doi:10.1016/j.promfg.2015.07.328.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 65-110). Westport, CT: Praeger.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 433-470). Westport, CT: Praeger.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7(1),* 33–50. https://doi.org/10.5281/zenodo.3554705

Hernandez M., Goldberg, B., Robson R., Owens, K., Blake-Plock, S., Welch, T., Ray, F. (2022) Enhancing the Total Learning Architecture for Experiential Learning. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Orlando, FL. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2022&AbID=112896&CID=944

IEEE (2022). Recommended Practice for Defining Competencies. IEEE 1484.20.2-2022. IEEE, Piscataway, NJ. https://standards.ieee.org/ieee/1484.20.2/10743/

Jia, B., Zhu, Z., & Gao, H. (2021). International comparative study of statistics learning trajectories based on PISA data on cognitive diagnostic models. *Frontiers in Psychology: Quantitative Psychology and Measurement.* https://doi.org/10.3389/fpsyg.2021.657858

Johnson, E., Lucas, G., Kim, P., Gratch, J. (2019). Intelligent Tutoring System for negotiation skills training. In Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds) Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science(), vol 11626. Springer, Cham. https://doi.org/10.1007/978-3-030-23207-8_23.

Johnson, M. S., Sinharay, S., & Bradlow, E. T. (2006). 17 Hierarchical item response theory models. Handbook of Statistics, 26, 587-606. https://doi.org/10.1016/S0169-7161(06)26017-6

Johnson, W. L., & Valente, A. (2009). Tactical language and culture training systems: Using AI to teach foreign languages and culture. *AI Magazine, 30(2),* 72. https://doi.org/10.1609/aimag.v30i2.2240

Johnson, W. L., Vilhjalmsson, H., & Samtani, P. (2005). The tactical language training system. http://secom.ru.is/publications/AIIDE2005Demo.pdf

Kay, J., & Lum, A. (2004). *Ontologies for scrutable student modelling in adaptive e-learning*. In Proceedings of the Adaptive Hypermedia and Adaptive Web-Based Systems Workshop on Semantic Web for E-Learning.

Kaya, Y., & Leite, W. L. (2017). Assessing Change in Latent Skills Across Time with Longitudinal Cognitive Diagnosis Modeling: An Evaluation of Model Performance. *Educational and Psychological Measurement, 77(3),* 369–388. https://doi.org/10.1177/0013164416659314.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: the four levels (Third Edition).* San Francisco: Berrett-Koehler.

Kizil, R. C. (2015). *The marginal edge of learning progressions and modeling: Investigating diagnostic inferences from learning progressions assessment* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (Accession No. 3743727).

Kohlberg, L. (1958). *The Development of Modes of Thinking and Choices in Years 10 to 16*. Ph. D. Dissertation, University of Chicago.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86(1),* 42–78. https://doi.org/10.3102/0034654315581420

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 387-432). Westport, CT: Praeger.

Lesgold, (2012). Practical issues in the deployment of new training technology. In P. J. Durlach & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (pp. 289-302). Cambridge University Press.

Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (1988). SHERLOCK: A coached practice environment for an electronics troubleshooting job. Accession Number ADA201748 Defense Technical Information System.

Levin, H. M. (2012). More than just test scores. *Prospects: Quarterly Review of Comparative Education, 42,* 269–284.

Levine, E. & Patrick, S. (2019). What is competency-based education? An updated definition. Vienna, VA: Aurora Institute. https://aurora-institute.org/wp-content/uploads/what-is-competency-based-education-an-updated-definition-web.pdf

Lorello, G. R., Cook, D. A., Johnson, R. L., & Brydges, R. (2014). Simulation-based training in anaesthesiology: A systematic review and meta-analysis. *British Journal of Anaesthesia, 112(2),* 231-245. https://doi.org/10.1093/bja/aet414

Ma, C., Ouyang, J., & Xu, G. (2022). *Learning Latent and Hierarchical Structures in Cognitive Diagnosis Models.* https://arxiv.org/pdf/2104.02143.pdf

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106(4*), 901–918. https://doi.org/10.1037/a0037123

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

McClarty, K. L., & Gaertner, M. N. (2015). Measuring mastery: Best practices for assessment in competency-based education. AEI Series on Competency-Based Higher Education. Washington, DC: American Enterprise Institute: Center on Higher Education Reform. https://files.eric.ed.gov/fulltext/ED557614.pdf

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. Academic Medicine, 65, S63–7. doi: 10.1097/00001888-199009000-00045.

Mor, Y. (2013). SNaP! Re-using, sharing and communicating designs and design knowledge using scenarios, narratives and patterns. In R. Luckin, S. Puntambekar, P. Goodyear, B. L. Grabowski, I. Underwood, J., & N. Winters (Eds). *Handbook of Design in Educational Technology* (pp. 189–200). London, UK: Routledge.

Mills, J.-A., Middleton, J. W., Schafer, A., Fitzpatrick, S., Short, S., & Cieza, A. (2020). Proposing a re-conceptualisation of competency framework terminology for health: a scoping review. *Human Resources for Health 18(1),* 1-16.

Nesbit, J. C., Adesope, O. O., Liu, Q., & Ma, W. (2014). How Effective are Intelligent Tutoring Systems in Computer Science Education? 2014 IEEE 14th International Conference on Advanced Learning Technologies, pp.99-103. DOI: https://doi.org/10.1109/ICALT.2014.38

Nickow, A., Oreopoulos, P., Quan, V. (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. *NBER Working Paper 27476* http://www.nber.org/papers/w27476.

O*NET OnLine (2022). https://www.onetonline.org/help/online/zones

Pham, D. (2019). Bringing learning back in: Examining three psychometric models for evaluating learning progression theories. Doctoral Dissertations. 1499. https://doi.org/10.7275/13487876 https://scholarworks.umass.edu/dissertations_2/1499

Piaget, J. (1952). The Origins of Intelligence in Children. New York, NY: W.W. Norton & Co. https://doi.org/10.1037/11494-000

Pokorny, B., Haynes, J., Gott, S., Chi, M. & Hegarty, M (2013). Infusing Simulations with Expert Mental Models, Adaptivity, and Engaging Instructional Interactions. *Proceedings from the Interservice/Industry Training, Simulation, and Education Conference*, Orlando FL.

Popham, W. J. (2001). Teaching to the test. *Educational Leadership, 58(6),* 16-20.

Reckase, M. D. (2009). Multidimensional item response theory. New York: Springer. DOI:10.1007/978-0-387-89976-3

Reich, J. & Huttner-Loan (2020). 0.502x Competency-based education: The why, what, and how. MIT Open Learning Library. https://openlearninglibrary.mit.edu/courses/course-v1:MITx+0.502x+1T2019/about

Rijmen, F., 2011. The latent class model as a measurement model for situational judgment tests. *Psychologica Belgica, 51(3-4)*, 197–212. DOI: http://doi.org/10.5334/pb-51-3-4-197

Robinson, C. (2018). Occupational Mobility, Occupation Distance, and Specific Human Capital, *The Journal of Human Resources 53(2),* 513-551.

Rosson, M. B., & Carroll, J. M. (2012). Scenario-based design. In J. A. Jacko (Ed.), *Human Computer Interaction Handbook, 3rd edition*. Boca Raton, FL: CRC Press. ISBN:9780429103971

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

Shippmann, J.S., Ash, R.A., Battista, M., Carr, L., Eyde, L.D., Hesketh, B., Keyhoe, J., Pearlman, K., Prien, E.P., & Sanchez, J.I. (2000). The practice of competency modeling. Personnel Psychology, 53, 703-740.

Shute, V., & Zapata-Rivera, D. (2012). Adaptive Educational Systems. In P. Durlach & A. Lesgold (Eds.), *Adaptive Technologies for Training and Education (pp. 7-27).* Cambridge: Cambridge University Press. doi:10.1017/CBO9781139049580.004

Silva, E., White, T., & Toch, T. (2015). *The Carnegie Unit: A century-old standard in a changing educational landscape.* Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from https://www.carnegiefoundation.org/wp-content/uploads/2015/01/Carnegie_Unit_Report.pdf on January 23, 2023.

Sinatra, A. M., Graesser, A. C., Hu, X., Goldberg, B., Hampton, A. J., & Johnston, J. H. (2022). *Design recommendations for intelligent tutoring systems, Volume 9: Competency-based scenario design.* Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center Simulation and Training Technology Center.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Harvard University Press.

SHRM (2017). Managing Organizational Change. https://www.shrm.org/resourcesandtools/tools-and-samples/toolkits/pages/managingorganizationalchange.aspx

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105(4),* 970–987. https://doi.org/10.1037/a0032447

Stevens, G. W. (2012). A critical review of the science and practice of competency modeling. *Human Resource Development Review, 12(1),* 86–107. https://doi.org/10.1177/1534484312456690

Steiner, E. D., Hamilton, L. S., Peet, E. D., & Pane, J. F. (2015). *Continued progress: Promising evidence on personalized learning: Survey results addendum.* Santa Monica, CA: RAND Corporation. Retrieved on November 12, 2022 from https://www.rand.org/pubs/research_reports/RR1365z2.html

Stodel, E. J., Wyand, A., Crooks, S., Moffett, S., Chiu, M., & Hudson, C. C. C. (2015). Designing and implementing a competency-based training program for anesthesiology residents at the University of Ottawa. *Anesthesiology Research and Practice,* 713038. https://doi.org/10.1155/2015/713038

Su, Y., Cheng, Z., Luo, P., Wu, J., Zhang, L., Liu, Q., Wang, S. (2021). Time-and-Concept Enhanced Deep Multidimensional Item Response Theory for interpretable Knowledge Tracing. *Knowledge-Based Systems, 218,* 106819. DOI: 10.1016/j.knosys.2021.106819

Su, Y., Cheng, Z., Dong, Y., Huang, Z, Wu, L., Chen, E., Wang, S., & Xie, F. (2022). Graph-based cognitive diagnosis for intelligent tutoring systems. *Knowledge-Based Systems, 253,* 109547.

Thomson, M. (2022, February 15). Synthetic Training Environment offers multi-dimensional combat preparation. Army Futures Command. https://www.army.mil/article/254005/synthetic_training_environment_offers_multi_dimensional_combat_preparation

Tossell, C., Wiese, E., Garritty, M. J., Denning, T., & Alliger, G. M. (2006). Developing Command and Control Performance-Based Training Criteria in a Network Centric Environment. Paper presented at the 11th International Command and Control Research and Technology Symposium (ICCRTS). Cambridge, U.K.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse (2010).

U.S. Department of Education (2016). *Non-regulatory guidance: Using evidence to strengthen education investments*. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

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

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

Western Governor's University (WGU) (2022). Retrieved 6/10/2022 from https://www.wgu.edu/

Xin, T., Wang, C., Chen, P., & Liu, Y. (2022). Editorial: Cognitive diagnostic models: Methods for practical applications. Frontiers in Psychology, Section Quantitative Psychology and Measurement. https://doi.org/10.3389/fpsyg.2022.895399

BIOGRAPHIES

Editors

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

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

Dr. Xiangen Hu is a professor in the Department of Psychology, Department of Electrical and Computer Engineering and Computer Science Department at The University of Memphis (UofM) and senior researcher at the Institute for Intelligent Systems (IIS) at the UofM and is professor and Dean of the School of Psychology at Central China Normal University (CCNU). Dr. Hu received his MS in applied mathematics from Huazhong University of Science and Technology, MA in social sciences and Ph.D. in Cognitive Sciences from the University of California, Irvine. Dr. Hu is the Director of Advanced Distributed Learning (ADL) Partnership Laboratory at the UofM, and is a senior researcher in the Chinese Ministry of Education’s Key Laboratory of Adolescent Cyberpsychology and Behavior.

Dr. Hu's primary research areas include Mathematical Psychology, Research Design and Statistics, and Cognitive Psychology. More specific research interests include General Processing Tree (GPT) models, categorical data analysis, knowledge representation, computerized tutoring, and advanced distributed learning.

Dr. Gregory Goodwin is a Team Leader and senior research scientist at the Combat Capabilities Development Command – Soldier Center in Orlando, FL. Prior to that, he worked in academia. After working in academia, Dr. Goodwin has spent the last fifteen years working for the Army researching ways to improve training methods and technologies. He holds a Ph.D. in Psychology from Binghamton University and an M.A. in Psychology from Wake Forest University.

Dr. Vasile Rus is a Professor in the Department of Computer Science at The University of Memphis with a joint appointment in the Institute for Intelligent Systems. Dr. Rus is also a Systems Testing Research Fellow of the Fedex Institute of Technology, a honor received for his pioneering work in the area of software systems testing. His research interests lie at the intersection of artificial intelligence, machine learning, and computational linguistics with an emphasis on developing interactive intelligent systems based on strong theoretical findings in order to solve critical challenges that would change the educational and human computer interaction landscape. Dr. Rus has been involved in research and development projects in the areas of computational linguistics and information retrieval for more than 15 years and in open-ended student answer assessment and intelligent tutoring systems for more than 10 years. He has been involved in the development of the following intelligent tutoring systems: DeepTutor (PI), Writing Pal (co-PI), MetaTutor (co-PI), and AutoMentor (co-PI). Dr. Rus has served in various roles on research projects funded by National Science Foundation, Department of Defense, and Department of Education, and private companies; has won the first two Question Answering competition organized by the National Institute for Science and Technology (NIST); recently his team won the English Semantic Similarity challenge organized by the leading forum on semantic evaluations – SemEval; has received 4 Best Paper Awards; produced more than 100 peer-reviewed publications; and currently serves as an Associate Editor of the International Journal on Tools with Artificial Intelligence and Program Committee member of the International Conference on Artificial Intelligence in Education (AIED 2015). Dr. Rus is 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).

Authors

Dr. Gautam Biswas conducts research in Intelligent Systems with primary interests in monitoring, control, and fault adaptivity of complex cyber physical systems. In particular, his research focuses on Deep Reinforcement Learning, Unsupervised and Semi-supervised Anomaly Detection methods, and Online Risk and Safety analysis applied to Air and Marine vehicle as well as Smart Buildings. His work, in conjunction with Honeywell Technical Center and NASA Ames led to the NASA 2011 Aeronautics Research Mission Directorate Technology and Innovation Group Award for Vehicle Level Reasoning System and Data Mining methods to improve aircraft diagnostic and prognostic systems.

Professor Biswas is also involved in developing intelligent open-ended learning environments focused on learning and instruction in STEM domains that adapt to students' learning performance and behaviors. He has also developed innovative learning analytics and data mining techniques for studying students' learning behaviors and linking them to their metacognitive and self-regulated learning strategies. His research is supported by funding from the Army, NASA, and NSF. He has published extensively, and currently has over 600 refereed publications. He is a Fellow of the IEEE Computer Society, Asia Pacific Society for Computers in Education, and the Prognostics and Health Management society.

Dr. Min Chi is an Associate Professor in the Department of Computer Science at North Carolina State University. Her research area lies in the interaction of Artificial Intelligence, machine learning, and human-computer interaction. She has established a foundational R&D portfolio with impactful advancements including Reinforcement Learning-based policy induction.

Fahmid Morshed Fahid is a fourth year PhD student in Computer Science at North Carolina State University, currently working as a graduate research assistant under the supervision of Dr. James Lester in the Center of Educational Informatics. His research interest involves student modeling and adaptive scaffolding using AI-driven educational learning environments. Fahmid served as a program committee member at AIED 2022. He completed his B.Sc in Computer Science from Bangladesh University of Engineering and Technology (2016) and his M.Sc in Computer Science from North Carolina State University (2022).

Dr. Peter Foltz is Research Professor at the University of Colorado's Institute of Cognitive Science and Executive Director of the National Science Foundation AI Institute for Student-AI Teaming. His work covers machine learning and natural language processing for educational and clinical assessments, large-scale data analytics, cognitive skills in reading and writing, team collaboration, and 21st Century skills learning, Much of his work has focused on NLP techniques for automatically analyzing the meaning of language through writing and speaking. The approaches are used for assessing abilities, for providing feedback, and for understanding underlying cognitive mechanisms in the brain. The methods he has pioneered are used by millions of people annually to improve achievement, expand student access, and make learning materials more affordable. He has served as the content lead for the framework development for Organisation of Economic Cooperation and Development's (OECD) Programme for International Student Assessment (PISA) assessments, including the 2018 Reading Literacy assessment, the 2015 assessment of Collaborative Problem Solving, and a new assessment of reading literacy for developing countries. He has been guest editor for a number of journals including *International Journal of AI in Education* and *Discourse Processes* as well as co-editor of the recent *Handbook of Automated Scoring: Theory into Practice.* His work has been covered widely in the press including The New York Times, Time Magazine, NPR, and Science. Peter has authored more than 150 journal articles, book chapters, and conference papers, as well as multiple patents. He previously worked at Pearson, New Mexico State University, Bell Communications Research, the Learning Research and Development Center at the University of Pittsburgh, Yale University, and the Harvard Institute for International Development.

Dr. Stephen B. Gilbert is currently Associate Director of Iowa State University's Virtual Reality Application Center and Director of its Human Computer Interaction graduate program. He is also an associate professor in the Industrial and Manufacturing Systems Engineering department. His research interests focus on technology to advance cognition, including intelligent tutoring systems, human-autonomy teaming, and XR usability. He works closely with industry, NSF, and DoD on research contracts and has also worked in commercial software development and runs his own company. He received a BSE from Princeton in civil engineering and operations research and a PhD from MIT in brain and cognitive sciences.

Dr. Benjamin Goldberg is a Senior Scientist at the U.S. Army CCDC Soldier Center, Simulation and Training Technology Center (STTC) in Orlando, FL. His research in Modeling & Simulation focuses on deliberate competency development, adaptive experiential learning in simulation-based environments, and how to leverage AI tools and methods to create personalized learning experiences. Currently, he is the lead scientist on a research program developing adaptive training solutions in support of competency development in Synthetic Training Environments. Dr. Goldberg is co-creator of the award winning Generalized Intelligent Framework for Tutoring (GIFT) and holds a PhD from the University of Central Florida.

Anisha Gupta is a doctoral student in Computer Science at North Carolina State University. Her research focuses on stealth assessment for student interactions with game-based learning environments.

Dr. Judy Kay is Professor of Computer Science in the Faculty of Engineering, University of Sydney. She heads the Human Centred Technology Research Cluster, a large multi-disciplinary research group that conducts fundamental research, design, engineering and evaluation of new technologies. She is a Payne-Scott Distinguished Professor at the University of Sydney, in recognition of her contributions both to education and to multi-disciplinary, high-impact and deployed research. A core focus of her research has been to create infrastructures and interfaces for personalisation, especially to support people in lifelong, life-wide learning. This ranges from formal education settings to supporting people in using their long-term ubicomp data to support self-monitoring, reflection and planning. Central to this has been the design of user modelling systems and interfaces. She has integrated these into new forms of interaction including virtual reality, surface computing, wearables and ambient displays. Her research has been commercialised and deployed and she has extensive publications in leading venues for research in user modelling, AIED, human computer interaction and ubicomp. She has held leadership roles in top conferences in these areas and is Editor-in-Chief of the IJAIED, International Journal of Artificial Intelligence in Education (IJAIED) and Editor of IMWUT, Interactive Mobile Wearable and Ubiquitous Technology (IMWUT).

Michael Krusmark is a Principal Research Scientist with CAE USA at the Air Force Research Laboratory, Human Performance Wing, Warfighter Interactions and Readiness Division. Mr. Krusmark holds a Bachelor of Science degree in Psychology (1990) and a Master of Arts degree in Cognitive Psychology (1997), both from Arizona State University. He possesses 20+ years of research experience on a wide range of projects aimed at developing and validating computational and mathematical models that replicate and extend the capacities of human cognition, and demonstrating the applicability of these capabilities in Air Force training domains. A primary focus of this work has been developing the Predictive Performance Optimizer, a patented technology for personalizing training that was co-invented by Mr. Krusmark.

Dr. Tiffany S. Jastrzembski is a Senior Cognitive Scientist with the Air Force Research Laboratory. She completed her undergraduate studies in cognitive psychology at Carnegie Mellon University, and attained her Masters and Doctoral degrees in the same field under the advisement of Dr. Neil Charness at the Florida State University. Her research focuses on the development of integrated cognitive and machine learning models capable of handling the dynamics of human memory, for purposes of delivering a precision learning capability for individual learners; namely, proficiency-based, optimized, personalized learning regimens. She has made noteworthy contributions in highly applied medical domains, linguist curricula at the Defense Language Institute, and total force training for all Air Force military and civilian personnel. She received the Air Force Research Laboratory's Early Career Award, the American Psychological Association's New Investigator Award, holds a patent on the Predictive Performance Optimizer software tool, and possesses a publication record of over 80 refereed papers.

Dr. Jong Kim is a research scientist at University of Central Florida, Orlando, FL at the time of the book publication. Dr. Kim received his PhD degree in Industrial Engineering at Pennsylvania State University, University Park, PA. His research interests lie in the area of cognitive science and engineering. Particularly, Dr. Kim is interested in theories of cognitive learning for the development of intelligent systems. Recently, Dr. Kim developed a theory of cognitive learning and unlearning (D2P: Declarative to Procedural) that is being applied to implement a series of intelligent training systems for the Navy in collaboration with Penn State and Charles River Analytics.

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. John's 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 & 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 21st 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.

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

Dr. James E. McCarthy. After completing his undergraduate work in Mathematics Education and Psychology, Dr. McCarthy moved on to Miami University where he earned his Masters and Doctorate degrees in Experimental Psychology. Following graduate studies at Miami University, Dr. McCarthy continued to develop his cognitive science skills as a research fellow at the Royal Victoria Hospital in Montreal. As Sonalysts' Vice President for Instructional System Development, Dr. McCarthy has nearly 30 years of experience in project management and training system development. He has devoted most of his career to the development and evaluation of advanced training systems for the DoD, other Governmental agencies, and commercial clients. Dr. McCarthy has also been instrumental in the development of a range of successful software products such as *ExpertTrain* – an engine that supports simulation-based intelligent tutoring, InTrain – an engine for adaptive interactive multimedia instruction, and RACE – an intelligent tutoring system authoring tool.

Jay Pande is a Ph.D. student in Computer Science at North Carolina State University. He received a B.S. in Computer Science from Duke University in 2020. His research interests lie in the use of speech data to improve educational outcomes for users of intelligent tutoring systems.

Dr. Bob Pokorny has a doctorate in Experimental Psychology from the University of Oregon in 1985; more than 15 years of experience in Department of Defense personnel research laboratories; and more than 20 years in private industry as a Human Factors Engineer, Instructional Designer, Knowledge Engineer and Director. His two primary roles at the Air Force Human Resources Laboratory were as Knowledge

Engineer for designing and implementing expert systems, and designing instruction and assessment to create and test the effectiveness of complicates Artificial Intelligence-based training systems. He has experience assessing program effectiveness, assessing training effectiveness, and designing assessments. He has experience in designing, implementing, and assessing simulations-based training. He has applied lessons from knowledge engineering and policy capture to construct computer-based systems to influence and assess behavior and performance.

Bob currently serves as Senior Cognitive Engineer at Affinity Associates. Affinity Associates applies human capital management concepts and best practices to personnel selection and classification, knowledge engineering, and human system design.

Dr. Steve Ritter is Founder and Chief Scientist at Carnegie Learning. He has been developing, analyzing and evaluating educational technology for over 20 years. He earned his Ph.D. in Cognitive Psychology at Carnegie Mellon University and was instrumental in the development and evaluation of the Cognitive Tutors for mathematics. He is the author of numerous papers on the design, architecture and evaluation of Intelligent Tutoring Systems and other advanced educational technology. He currently leads the research team at Carnegie Learning, focusing on improving the educational effectiveness of its products and services. Each year, over 500,000 students use Carnegie Learning's mathematics curricula.

Dr. Robert Sottilare is the Science Director for Intelligent Training at Soar Technology, Inc. He came to SoarTech in 2018 after completing a 35-year federal career in both Army and Navy training science and technology organizations. At the US Army Research Laboratory, he led the adaptive training science and technology program where the focus of his research was automated authoring, instructional management, and analysis tools and methods for intelligent tutoring systems (ITSs) and standards for adaptive instructional systems. He is the father of the Generalized Intelligent Framework for Tutoring (GIFT), an award-winning open source, AI-based adaptive instructional architecture. GIFT has over 2000 users in 76 countries.

Dr. Sottilare has long history as a leader, speaker, and supporter of learning and training sciences forums at the Defense & Homeland Security Simulation, HCII Augmented Cognition, and AI in Education conferences. He is the founding chair of the HCII Adaptive Instructional Systems (AIS) Conference. He is a member of the AI in Education Society, the Florida AI Research Society, the IEEE Computer Society and Standards Association (senior member), the National Defense Industry Association (lifetime member), and the National Training Systems Association. He is currently the IEEE Project 2247 working group chair for the development of standards and recommended practices for AISs. He is a faculty scholar and has been an adjunct professor at the University of Central Florida where he taught a graduate level course in ITS theory and design.

Dr. Kurt VanLehn is the Diane and Gary Tooker Chair for Effective Education in Science, Technology, Engineering and Math at Arizona State University. He is also a Professor of Computer Science. He received a Ph. D. from MIT in 1983 in Computer Science, and worked at BBN, Xerox PARC, CMU and the LRDC (University of Pittsburgh). He founded and co-directed two large NSF research centers (Circle; the Pittsburgh Science of Learning Center). He has published over 185 peer-reviewed publications. He is a fellow in the Cognitive Science Society and associate editor of the *International Journal of Artificial Intelligence in Education*. Dr. VanLehn has been working in the field of intelligent tutoring systems since such systems were first invented. Most of his current work explores new applications of this well-established technology. For example, FACT is an intelligent classroom orchestration system for helping middle school math teachers both deeply analyze student work and manage the flow of ideas and student work across individual, group and whole-class activities.

Dr. Diego Zapata-Rivera is Distinguished Presidential Appointee and director of the Learning and Assessment Foundations and Innovations (LAFI) Research Center at Educational Testing Service in Princeton, NJ. He earned a Ph.D. in computer science (with a focus on artificial intelligence in education) from the University of Saskatchewan in 2003.

His research at ETS has focused on the areas of innovations in score reporting and technology-enhanced assessment including work on adaptive assessment and learning environments, conversation-based assessment, caring assessment, and game-based assessment. His research interests also include Bayesian student modeling, open student models, virtual environments, authoring tools and program evaluation.

Dr. Zapata-Rivera has produced over 140 publications including edited volumes, journal articles, book chapters, and technical papers. Dr. Zapata-Rivera is an elected member of the International AI in Education Society Executive Committee (2022-2027), a member of the Editorial Board of User Modeling and User-Adapted Interaction, and an Associate Editor for AI for Human Learning and Behavior Change. Dr. Zapata-Rivera has contributed his expertise to projects sponsored by the National Research Council, the National Science Foundation, NASA and the US Army Research Laboratory.

Design Recommendations for Intelligent Tutoring Systems

Volume 10
Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analysis of Intelligent Tutoring Systems

Design Recommendations for Intelligent Tutoring Systems (ITSs) explores the impact of intelligent tutoring system design on education and training. Specifically, this volume includes Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analyses of Intelligent Tutoring System components. It includes overview chapters and individual chapters which examine the current and potential future states of both traditional and advanced intelligent tutoring system elements. These topic areas include Learner Modeling, Instructional Strategies, Authoring Tools, Domain Modeling, Assessment, Team Tutoring, Self-Improving Systems, Data Visualization, Competency-Based Scenario Design, and the Generalized Intelligent Framework for Tutoring (GIFT).

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 has been a member of the Generalized Intelligent Framework for Tutoring (GIFT) team since 2012.*
- ***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. Gregory Goodwin*** *is a Team Leader and Senior Research Scientist at U.S. Army Combat Capabilities Development Command – Soldier Center – Simulation and Training Technology Center.*
- ***Dr. Vasile Rus*** *is a Professor in the Department of Computer Science at the University of Memphis with a joint appointment in the Institute for Intelligent Systems.*

A Volume in the Adaptive Tutoring Series

9 780997 725834