A Modular Framework to Support the Authoring and Assessment of Adaptive Computer-Based Tutoring Systems (CBTS)

Robert A. Sottilare, Benjamin S. Goldberg, Keith W. Brawner and Heather K. Holden
U.S. Army Research Laboratory – Human Research and Engineering Directorate
Orlando, Florida
{robert.sottilare; benjamin.s.goldberg; keith.brawner; heather.k.holden} @us.army.mil

ABSTRACT

An emphasis on self-development in the military community has highlighted the need for adaptive computer-based tutoring systems (CBTS) to support point-of-need training in environments where human tutors are either unavailable or impractical. Effective human tutors ask questions, tailor feedback, provide opportunities for reflection, and change the content, direction, pace, and challenge level of instruction to optimize learning (e.g., acquisition of knowledge or skills). Adaptive CBTS also attempt to select optimal instructional strategies to meet the specific learning needs of individuals or teams. To make these optimal instructional strategy decisions, the adaptive CBTS assesses trainee attributes (e.g., progress, behaviors or physiology), uses these attributes to classify states and predict learning outcomes (e.g., performance, skill acquisition, retention), and then adapts the instruction to influence learning. A truly adaptive CBTS must have a suitable trainee model, a repertoire of instructional strategies, and a methodology for selecting the best strategy. Significant challenges in the design and development of adaptive CBTS include methodologies to: assess the influence of trainee attributes that inform positive/ negative learning states (e.g., confusion, boredom, frustration, and pleasure); and assess the influence of specific instructional strategies on learning given the learner's state and the training context (e.g., tasks, conditions, and learning objectives). This paper considers a modular tutoring system framework to support the authoring and assessment of adaptive tutoring capabilities. The Generalized Intelligent Framework for Tutoring (GIFT) supports authoring standards and allows users to manipulate models, libraries, and domain-specific content to empirically determine the influence of variables of interest (e.g., learning style, sensor data, feedback modes, and stress) on learning. The framework supports a variety of experimental views, including ablative tutor studies, tutor vs. traditional classroom training comparisons; evaluation of intervention vs. non-intervention strategies; pedagogical model comparisons; and tutor vs. tutor comparisons.

ABOUT THE AUTHORS

Robert A. Sottilare, Ph.D. is the Associate Director for Science & Technology within the U.S. Army Research Laboratory - Human Research and Engineering Directorate (ARL-HRED) and directs research within the Learning in Intelligent Tutoring Environments (LITE) Laboratory at ARL's SFC Paul Ray Smith Simulation & Training Technology Center (STTC). He has 28 years of experience in both U.S. Army and Navy training and simulation as a researcher, engineer and program manager. He leads the international program at STTC and chairs training technology panels within The Technical Cooperation Program (TTCP) and NATO. Dr. Sottilare holds a patent for a high-resolution, head-mounted projection display and his recent publications have appeared in the Educational Technology Journal, the Journal for Defense Modeling and Simulation and the proceedings of the Intelligent Tutoring Systems Conference 2010. He is a graduate of the Advanced Program Managers Course at the Defense Systems Management College, and his doctorate in modeling & simulation with a focus in intelligent systems is from the University of Central Florida. In January 2012, Dr. Sottilare was honored as the inaugural recipient of the U.S. Army Research Development & Engineering Command's (RDECOM's) Modeling & Simulation Lifetime Achievement Award. The focus of his current research is on the application of artificial intelligence tools and methods to adaptive training environments.

Benjamin S. Goldberg is a member of the Learning in Intelligent Tutoring Environments (LITE) Lab at the U.S. Army Research Laboratory's (ARL) Simulation and Training Technology Center (STTC) in Orlando, FL. He has been conducting research in the Modeling and Simulation community for the past 3 years with a focus on adaptive learning and how to leverage Artificial Intelligence tools and methods for adaptive computer-based instruction. Mr. Goldberg is a Ph.D. student at the University of Central Florida and holds an M.S. in Modeling and Simulation. Prior to employment with ARL, he held a Graduate Research Assistant position for two years in the Applied

Cognition and Training in Immersive Virtual Environments (ACTIVE) Lab at the Institute for Simulation and Training. Recent publications include a proceedings paper in the 2010 International Conference on Applied Human Factors and Ergonomics and a poster paper presented at the 2010 Spring Simulation Multiconference.

Keith W. Brawner is a researcher for the Learning in Intelligent Tutoring Environments (LITE) Lab within the U. S. Army Research Laboratory – Human Research and Engineering Directorate (ARL-HRED). He has 4 years of experience within U.S. Army and Navy acquisition, development, and research agencies. He holds a Masters degree in Computer Engineering with a focus on Intelligent Systems and Machine Learning from the University of Central Florida, and is currently in pursuit of a doctoral degree in the same field. The focus of his current research is in machine learning, adaptive training, trainee modeling, and knowledge transfer.

Heather K. Holden, Ph.D. is a researcher in the Learning in Intelligent Tutoring Environments (LITE) Lab within the U.S. Army Research Laboratory – Human Research and Engineering Directorate (ARL-HRED). The focus of her research is in artificial intelligence and its application to education and training; technology acceptance and Human-Computer Interaction. Dr. Holden's doctoral research evaluated the relationship between teachers' technology acceptance and usage behaviors to better understand the perceived usability and utilization of job-related technologies. Her work has been published in the Journal of Research on Technology in Education, the International Journal of Mobile Learning and Organization, the Interactive Technology and Smart Education Journal, and several relevant conference proceedings. Her PhD and MS were earned in Information Systems from the University of Maryland, Baltimore County. Dr. Holden also possesses a BS in Computer Science from the University of Maryland, Eastern Shore.

A Modular Framework to Support the Authoring and Assessment of Adaptive Computer-Based Tutoring Systems (CBTS)

Robert A. Sottilare, Benjamin S. Goldberg, Keith W. Brawner and Heather K. Holden
U.S. Army Research Laboratory – Human Research and Engineering Directorate
Orlando, Florida
{robert.sottilare; benjamin.s.goldberg; keith.brawner; heather.k.holden} @us.army.mil

INTRODUCTION

The Army Learning Concept (ALC) for 2015 (TRADOC, 2011) has defined a need to "develop adaptive, thinking Soldiers and leaders capable of meeting the challenges of operational adaptability in an era of persistent conflict." To meet these challenges, the Army has placed additional emphasis on the self-development to augment institutional training. Soldiers will be largely responsible for their learning, but will clearly need guidance. ALC for 2015 describes a career-long, individually tailored Army Learning Model (ALM) that is expected to support technology-based instruction adapted to the learner's competence and state (e.g., cognitive load, engagement, or motivational level) in a variety of training domains. Learners will experience instruction through artificially intelligent computer-based tutoring systems (CBTS) that support their development in the absence of human instructors.

If CBTS are expected to efficiently and effectively guide self-development, they will need to be on par with or be better than expert human tutors in assessing learner state (e.g., potential and performance), and selecting instruction strategies to optimize motivation and engagement to influence learning. Effective human tutors ask questions, tailor feedback, provide opportunities for reflection, and change the content, direction, pace, and challenge level of instruction to optimize the learner's learning (e.g., acquisition of knowledge or skills). To make the best possible instructional decisions, an adaptive CBTS must have a suitable learner model, a repertoire of instructional strategies and a methodology for selecting the best strategy. If each learner has differing needs, how will the CBTS optimize instructional decisions? Methods are needed to assess the influence of individual states and traits in learner models, and CBTS instructional strategies and tactics to ascertain their relationship to positive learning outcomes (e.g. skill development, knowledge acquisition, and retention).

Significant challenges in the design and development of adaptive CBTS include methodologies to: 1) assess the influence of learner attributes that inform positive and negative learning states (e.g., confusion, boredom, frustration, and pleasure) and 2) assess the influence of specific instructional strategies on learning given the cognitive and affective state of the learner(s) and the training context (e.g., tasks, conditions, and learning objectives).

This paper considers a modular tutoring system framework to support the assessment of adaptive tutoring capabilities. The Generalized Intelligent Framework for Tutoring (GIFT) provides three primary services for learners, instructional system designers, expert behavior modelers, training system developers, trainers, and researchers (see Figure 1): authoring of CBTS and CBTS components, tools and methods; management of instructional processes using best pedagogical practices based on the behaviors of expert human tutors; and an assessment methodology to evaluate the effectiveness of CBTS and CBTS components, tools and methods.

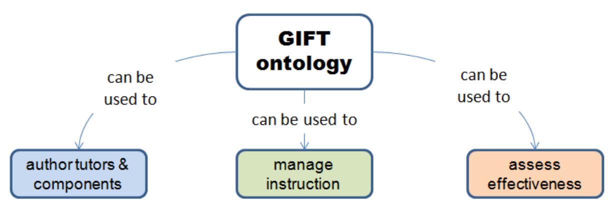

Figure 1: Primary services in GIFT

Authoring tools support the development and configuration management of user models and interfaces, domain-specific knowledge, instructional strategies, and CBTS compilers. Instructional process tools support the management of all phases of instruction, including initialization (including user authentication, pre-instruction (including surveys and mission briefings), instruction (including supervision of the tutoring process, instructional strategy selection, and tutoring tactics) and post instruction (including surveys and after-action reviews). Assessment tools evaluate learner performance, learning gain, and skill development. These tools also evaluate the tutor's performance and learning effect of components, tools, and methods.

UTILITY OF A CBTS FRAMEWORK

While the target beneficiary of a CBTS framework is the learner who can receive tutoring tailored to his/her needs and capabilities, let us examine how other users might exploit standard tools and methods in a generalized CBTS framework like GIFT. As mentioned in the introduction, in addition to learners, GIFT users include instructional system designers, expert behavioral modelers, training system developers, trainers, and researchers. User interfaces must be unique to each of these disciplines to support their authoring and assessment activities.

Instructional system designers (ISDs) incorporate "known and verified learning strategies into instructional experiences which make the acquisition of knowledge and skill more efficient, effective, and appealing" (Merrill, Drake, Lacy, Pratt, and ID2 Research Group, 1996). Their tasks include elements of both authoring and assessment of instructional objects (e.g., scenarios, instructional strategies, instructional content, and performance assessment tools). ISDs can use GIFT to author strategies aligned with a particular instructional theory (e.g., Gagne's Theory of Instruction; Gagné, 1985; Gagné & Driscoll, 1988). They may also assess instructional strategies applied to self-regulated computer-based learning environments using instructional design models (e.g., Analyze, Design, Develop, Implement and Evaluate (ADDIE); Branson, Rayner, Cox, Furman, King, and Hannum, 1975; Kemp, Morrison, and Ross; 1994).

Developers of behavioral models are interested in capturing behaviors (e.g., decisions, actions) that define expertise for a particular task or set of tasks. Their mission is to evaluate essential behaviors to author expert models. They can use GIFT to map the successful paths of experts through courses of instruction to develop expert/ideal models. The same tactic can also be used with novices to identify common errors or misconceptions, and successful strategies for overcoming these errors/misconceptions. These models are used for performance assessment purposes and link learner behavior with defined task objectives.

Training system developers take a more holistic approach to adapting instructional content to a selected training environment. In military training systems significant emphasis is placed on after-action reviews and much less emphasis is placed on near real-time feedback to the learner. Timely feedback is generally more important to domain novices than it is to more experienced and proficient learners. Training system developers using GIFT have access to libraries of strategies that are tailored to the user and can be used to develop timely feedback mechanisms. Tools like the GIFT survey tool are employed to promote reuse of survey instruments used to elicit information and test learner knowledge. GIFT enables training developers to account for individual differences associated with the learner population for the purpose of tailoring content and providing guidance specific to learner needs.

In Figure 2, GIFT modules interact and receive/pass information through a Service-Oriented Architecture (SOA). Part of the information communicated is the domain knowledge which is comprised of tasks, conditions, standards, media content, instructional strategies, questions, misconceptions, and other information specific to that learning domain (e.g., land navigation). Training system developers have the option of integrating with GIFT through the SOA to provide the domain knowledge that GIFT uses to make pedagogical decisions and implement instructional strategies and tactics, or they can choose to embed the domain independent pedagogical strategies/tactics within their training system.

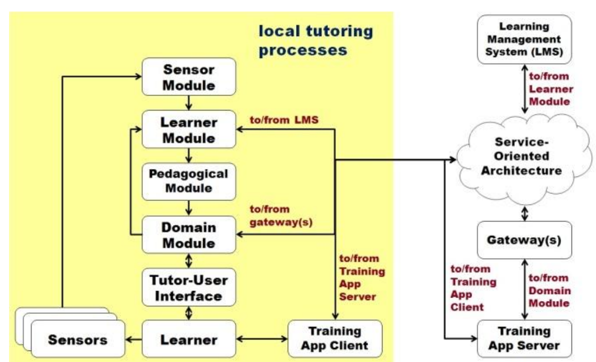

Figure 2: Training system integration with GIFT

Finally, GIFT allows researchers to manipulate the construct of the learner model, instructional strategy library, and/or domain-specific content to empirically determine the influence of variables of interest (e.g., learner states and traits) on learning outcomes. The framework supports a variety of experimental views, including ablative tutor studies, tutor vs. traditional classroom training comparisons; evaluation of intervention vs. non-intervention strategies;

pedagogical model comparisons; and tutor vs. tutor comparisons.

In the following sections, we review the construct and implications of learner models, instructional strategies, and assessment tools for learner performance and tutor effect within GIFT. The discussion provides practical implications for using a framework such as GIFT to support CBTS authoring and instructional delivery to large distributed organizations. In this context, the authors also put forward recommendations for future research and development.

CBTS LEARNER MODELING

The learner model is what the CBTS uses to make instructional decisions before and during tutoring sessions. A major research question for learner modeling is "what should be in a learner model and how is this information derived?" Since we want to use the learner's states and traits to make optimal instructional decisions for learning, a logical place to begin answering this question is to understand different learning domains associated with broad sets of tasks presented by CBTS. In the context of understanding learning to support capabilities within a CBTS, GIFT was focused on four types of learning: cognitive (thinking), affective (feeling), psychomotor (doing), and social learning (collaborating).

Cognitive learning is demonstrated by behaviors indicating increasingly complex and abstract mental capabilities. Ranging from least complex to most complex, these behaviors include remembering, understanding, applying, analyzing, evaluating, and creating (Anderson and Krathwohl, 2000). To exercise remembering, a CBTS could present a problem where the learner is asked to define or list terms (e.g., define bounding overwatch or list the inert elements in the periodic table).

Assessing simple concepts like remembering can be straightforward for expert human tutors who easily recognize learner behaviors. A CBTS could use Latent Semantic Analysis (LSA) to compare the learner's response to the correct answer in the domain model, but assessing even this simple concept can quickly become complicated. How might the CBTS assess and react to a lengthy delay by the learner in providing a response? Does this mean that the learner struggled to recall the information, withdrew from the process, or was distracted? In addition to the response, the CBTS might need other information (e.g., response time, affective state) about the learner to understand the learner's state and inform its instructional decisions in cognitive domains. It may also be difficult to separate affect from cognitive learning. Graesser and D'Mello (2012) posit that learners encounter a state of cognitive disequilibrium when confronting difficulties. Cognitive-affective processes interact until equilibrium is restored. The opposite may also be true in that emotion may limit cognition and cause disequilibrium.

"The extent to which emotional upsets can interfere with mental life is no news to teachers. Students who are anxious, angry, or depressed don't learn; people who are caught in these states do not take in information efficiently or deal with it well" (Goleman, 1995).

In affective domains, learning is demonstrated through behaviors indicating emotional growth and maturity, including, in ascending order of complexity, receiving, responding, valuing, organizing, and characterizing by values (Krathwohl, Bloom, and Masia, 1964). To exercise skills related to "characterizing by values", a CBTS could present a problem where the learner is asked to make a moral judgment (e.g., loyalty versus honesty). A task like this may be complicated by conflicts between individual and organizational values resulting in emotional states that are counterproductive to learning objectives. The CBTS would need knowledge of the learner's value system as well as the organizational values to assess the learner's behaviors and select appropriate interventions to promote affective growth also known as emotional intelligence (Goleman, 1995). Affective learning also influences how individuals manage cognitive resources. Applying strategies to enhance motivation to learn instills behaviors of perseverance and enthusiasm to continue when challenge is present. In addition, research shows instructor immediacy, where verbal and non-verbal feedback is provided during task interaction, acts as an affective mediator during cognitive learning tasks (Rodríguez, Plax, & Kearney, 1996).

In psychomotor learning domains, the emphasis is on the relationship between the learner's cognitive functions and their physical skills (e.g., coordination, strength, or speed). Per Simpson (1972), psychomotor learning behaviors include perceptions (awareness), sets (readiness), guided responses (attempts), mechanisms (basic proficiency), complex overt responses (expert proficiency), adaptation (adaptive proficiency), and origination (creative proficiency).

To exercise marksmanship skills, for instance, a CBTS could present a problem where the learner is asked to detect, identify, and prosecute intended targets. During this psychomotor task, the CBTS could collect behavioral data on the learner's breathing, trigger pressure, weapon pitch, weapon cant, and weapon aim point, and relate this data to successful/unsuccessful performance outcomes based on either the individual learner's historical performance, or common successful behaviors/errors experienced by the population.

While most CBTS have emphasized individual learning, military training is generally focused on how Warfighters support team performance objectives, and social learning plays a large role in the development of military skills. To exercise social learning skills, a CBTS could present a problem where a unit is asked to plan, rehearse, and execute a building clearing mission. Members of the unit likely vary in experience, task competence, and motivational level. Individual mission roles may not be identical. The members of the unit may or may not have worked together previously. Trust level and communications may be affected by leadership, clarity of the mission objectives, individual roles/responsibilities, and task interdependence. It may be necessary to have team models of performance, competence, trust, cognition, affect, and communications for the CBTS to ascertain the healthy functioning of the team (Sottilare, 2011).

While the multi-learner modeling necessary for team training with CBTS make assessments more complex, the opportunities for positive learning through social interaction make the return worth the investment in researching and developing these capabilities. Peer tutoring within the academic domains have been found to significantly deepen the learning gains for both the tutor (Sharpley, Irvine, and Sharpley, 1983; Robinson, Schofield, and Steers-Wentzell, 2005) and the learner (Ogan, Finkelstein, Walker, Carlson, and Cassell, 2012). Such environments foster social motivation (Rohrbeck, Ginsburg-Block, Fantuzzo, and Miller, 2003) and increase engagement (Webb, 1989).

Unobtrusive physiological and behavioral sensors and associated algorithms to process and classify learner states are needed to support adaptive, tailored tutoring. Learner modeling should include data capable of informing instructional strategy selection that, in turn, influences learning. Given the lack of standards for learner modeling and the varied opinions in the literature regarding learner modeling structure, the GIFT messaging construct has been designed to support modular models. The implications of this design are that a variety of learner models can be evaluated in the same testbed using GIFT as long as they conform to the messaging protocol, a JavaScript Object Notation (JSON), in a programming language-independent format.

In summary, learner modeling may be influenced in part by the domain tasks being trained. The learning domain (e.g., cognitive, affective, or psychomotor) also influences the variables of interest that might be part of the learner model structure. Problems presented by a CBTS may include elements of one or more learning domains, and the learner model should be structured to collect/store relevant data to ascertain the current and future progress of the learner. Their cognitive states (e.g., workload and engagement), affective states (e.g., emotional and motivational states), psychomotor performance, and team performance should be available from the learner model to support instructional decisions.

CBTS INSTRUCTIONAL STRATEGIES

Following the discussion of learner models, this section discusses their relationship to instructional strategy selection for adaptive and tailored tutoring. GIFT has been designed to support strategic (higher level decision making) and associated subordinate tactical level interventions that vary by method (how). This is necessary if the framework is to remain domain-agnostic. Within GIFT, instructional strategies include two primary approaches, macro-adaptation and micro-adaptation that may be used separately, but are most effective when used together with CBTS.

Macro-adaptation is a strategy that uses the learner model states (e.g., affective) and traits (e.g., interests) to initialize the domain model prior to instruction. Essentially, macro-adaptation structures and organizes a set of tactics to be implemented during training and addresses four design areas: (1) selection, (2) sequencing, (3) synthesizing, and (4) summarizing (Reigeluth, 1999). This approach uses information known about the learner prior to system interaction to apply off-line tailoring of content and guidance mechanisms. The macro-adaptive strategy may include tactics for selecting the starting point for instruction and challenge level of the scenario based on the learner's domain competency, the selection of media based on individual preferences (e.g., personality factors), options for the degree of control assumed by the learner during instruction based on domain competency, and even the selection of a tutor proxy (e.g., embodied conversational agent) based on learner preferences. Macro-adaptation provides a starting point for instruction based on what the CBTS knows about the learner.

Micro-adaptive strategies include near real-time tactics for adapting instruction to meet learning needs based on the learner's behaviors during instruction. Scaffolding is a micro-adaptive tactic where the CBTS provides less and less support as the competency of the learner grows. Another micro-adaptive tactic is feedback management where feedback (type, frequency, delivery method) is selected based on the learner's current and/or projected future state. These types of interventions are triggered by state determinations within the learner model, and applied micro-adaptive strategies must be tied to the specific problem context the learner is interacting within. The approach uses task performance and state variables to determine progress, reaction to training, and executes adaptation in real-time (Park & Lee, 2004).

Macro and micro-adaptive strategies can be compared to betting on a horse race. Macro-adaptive tactics are like betting on a horse before the race starts based on the past history, trainer reports, and other data about the horses. Micro-adaptive tactics are like betting on the horse at each stage of the race as it progresses based on its current and projected state (e.g., potential). It is good to know about the horse before the race, but there is a much better chance of being able to influence the outcome by introducing effective tactics during the race. The implications for tutoring using macro and micro-adaptive strategies include: knowledge of the bigger picture via historical data and trends of analysis; and agility during tutoring to make changes when learning momentum shifts. These strategies make it easier to provide continuity from one tutoring session to the next, and to identify objectives and trends essential to assessing learner performance.

ASSESSING LEARNER PERFORMANCE

GIFT builds on the U.S. Army construct of tasks, conditions, and standards to assess performance. Tasks (or problems) are associated with concepts to be learned. Concepts are bundled to form lessons. Multiple lessons form a course of learning. Thereby, multiple tasks (separate, concurrent, or overlapping) are presented for the learner to execute under a set of conditions (e.g., in a building with hostile combatants) and assessed against a set of standards, which are part of the domain knowledge in the CBTS. Tasks can be classified as well-defined or ill-defined, and simple or complex. Well-defined tasks generally have clearly defined outcomes and standards. Ill-defined tasks may have many pathways to success with more loosely defined standards (e.g., constraint-based). Well-defined tasks may be complex in that they are composed of a long and winding, but clearly defined process. In addition, well-defined tasks may be executed in variable conditions and environments that influence how actions are executed, requiring adaptability on the part of the learner. Ill-defined tasks may be simple in that they have few transitions from one key concept to another. It is self-evident that assessing real-time performance for ill-defined, complex tasks is most difficult due to the lack of clear standards and the multi-faceted nature of the tasks. Performance assessment methods evaluated for use in GIFT range from simple heuristics based on clear standards to classification algorithms to Partially-Observable Markov Decision Processes (POMDPs).

POMDPs are well-suited for assessing performance in domains where uncertainty is high. POMDPs use current states from the learner model to determine which actions and resulting transitions (to the same or other states) will result in the highest reward (successful outcome). By projecting forward, POMDPs can assess multiple outcomes to determine the most efficient and effective paths to success. A drawback to using POMDPs is when the number of states in the learner model considered to project an outcome is very large. Carefully crafting the learner model to contain only the most influential variables for learning may reduce the uncertainty associated with assessing performance for ill-defined tasks when using POMDPs.

Regardless of the assessment method, GIFT categorizes performance as "meeting expectations", "below expectations", or "above expectations". These assessments are used for triggering defined micro-adaptive tactics when interventions are deemed appropriate. Each of these categories invokes different instructional tactics based on the context of the instruction, the learner's performance trends (local and global), and the learner's states. The problem of assessing performance is compounded when we begin to evaluate teams.

A team's learning potential is maximized when each individual actively participates in the learning task, thereby, increasing the probability that all trainees understand the learning material and no one is left behind (Soller, 2001). Team performance, however, is dependent on more than just the sum of individual performances. Depending on the task, conditions, and standards, Sottilare, Holden, Brawner and Goldberg (2011) note that the assessment of team performance is likely to include assessment of the interdependency of the roles and responsibilities of the team members, leadership roles and communication, understanding of roles by the team, the domain competency of team members, trust within the team (credibility and reliability), and finally, collective models of team cognition (e.g., shared mental models, workload, and engagement) and affect (e.g., emotions, motivation). In turn, each of these team states may be difficult to assess and, thereby, limit the adaptability of the CBTS.

Now that we have reviewed individual and team performance assessments, we are ready to discuss the assessment of learning effect in CBTS. From CBTS researcher and developer perspectives, understanding the impact (effect size) of various tutoring methods and structures is essential in optimizing CBTS performance and, thereby, learning.

ASSESSING LEARNING EFFECT

Learning effect (also known as learning gain) depends upon not only the utility of whole tutoring systems, but also the effectiveness of specific CBTS components, models, algorithms, methods, and instructional strategies/tactics. Learning effect has been a subject of debate for as long as there have been CBTS. Bloom (1984) compared traditional classroom training to one-to-one human tutoring. This is relevant since the instructional niche filled by most CBTS today is limited to one CBTS guiding one learner through a program of instruction. Bloom reported a resulting two (2) sigma effect size. This means that the average performance for one-to-one tutoring was found to be two standard deviations higher than the average performance for traditional classroom instruction. This is credited to one-on-one tutoring optimizing the time of instruction by focusing on the strengths and weaknesses associated with a given learner.

CBTS researchers have strived to realize similar effect sizes, but have fallen short. Effect sizes have been reported ranging from 0.42 sigma for unskilled human tutors (Cohen, Kulik, and Kulik, 1982) up to an average of about 1 sigma for CBTS, including Pittsburgh Urban Mathematics Project/PUMP Algebra Tutor (PUMP/PAT - Koedinger, Anderson, Hadley, and Mark, 1997); Andes Tutor (VanLehn, et al, 2005); and SHERLOCK (Lesgold, Lajoie, Bunzo, and Eggan 1988). Woolf (2011) reported similar results with achievement scores (performance) for CBTS instruction showing an average 1.05 sigma over traditional classroom methods. This result equates to an increased median achievement score of 85% versus 50% for traditional classroom instruction.

While the literature supports and confirms the learning effect of tutoring systems, it is also essential to be able to assess, compare and contrast the learning effect of tutoring technologies (components, tools, and methods) in a controlled environment using scientific methods. A testbed methodology (Figure 3), adapted from Hanks, Pollack and Cohen's (1993) controlled experimentation recommendations, has been designed within GIFT to support an experimental environment for a variety of tutoring studies.

The GIFT experimental methodology or testbed allows for interchangeable modular components, including learner models, instructional engines, and domain-specific knowledge. Similar to analyzing system-level effectiveness through the measurement of learning with and without the system, one can measure the effectiveness of individual components through testing the effect of their presence and absence. GIFT supports comparison studies, strategy and tactic effectiveness assessments, and ablative studies. These assessments can be used to drive future tutoring system design by highlighting the importance of individual differences (traits) in learner models, cognitive and affective state classification, and instructional strategy selection. The interchangeable modular design of GIFT supports assessment methodologies by facilitating the configuration of test cases.

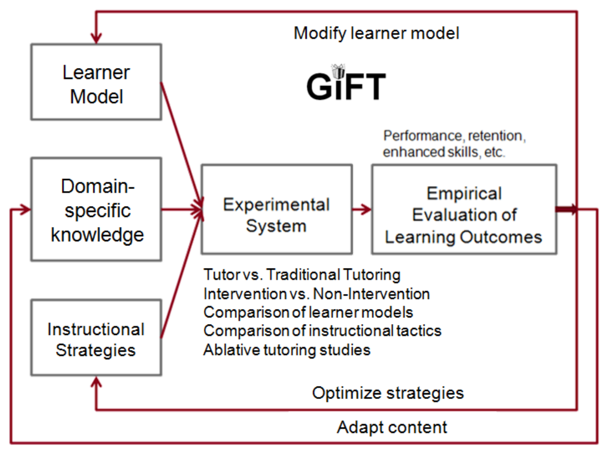

Figure 3: GIFT Experimental Methodology

The testbed may also be used to compare automated methods for domain knowledge (e.g., expert models) to more time-consuming methods to see if automated methods can produce the same effect sizes as manual methods. The experimental methodology supports spiral development of GIFT through iterative design improvements based on empirical evidence, and ultimately resulting in the 'platinum' tutoring capabilities (e.g., more effective than a human tutor) defined by Sottilare and Gilbert (2011) as a standard of performance for future CBTS.

DISCUSSION

Our discussion emphasizes how a standard framework like GIFT enables users (e.g., learners, designers, training developers, and researchers) to optimize system performance over time. Tailored learning, usability, traceability of research, and extensibility are highlighted.

Tailored Learning Validation Studies

Significant challenges persist in developing efficient and effective tutoring solutions that are tailored to the needs of individual learners. An essential element in supporting tailored learning experiences is the development and maintenance of an optimized learner model that is populated and updated with timely learner state and trait data to support instructional decisions.

As part of the development of the GIFT learner model, the Army Research Laboratory (ARL) has been conducting a series of experiments related to assessment of sensor data (Goldberg & Brawner, 2012; Carroll, et al, 2011; Brawner & Goldberg, 2012); assessment of learner state changes (Brawner & Gonzalez, 2011); macro-adaptation (pre-training tailoring) based on the learner's previous performance (Zook, Riedl, Holden, Sottilare, and Brawner, 2012); and learner perception of the tutor (Holden, 2012; Holden and Goldberg, 2011).

Experiments conducted under these efforts resulted in the following findings that are included in the GIFT development strategy. Carroll, et al (2011) evaluated physiological and behavior sensors to determine the efficacy of assessing cognitive and affective states. Key cognitive states for learning were identified as engagement, attention, and workload. Engagement was classified using unobtrusive chair pressure sensors and a low-cost electro-encephalograph (EEG). Attention and workload were classified using a low-cost EEG and eye tracker. Key affective states for learning were identified as anger/frustration, fear/anxiety, and boredom. Anger/frustration and boredom were classified using a motion detector, heart rate monitor, and/or chair pressure sensors. Fear/anxiety was classified using the heart rate monitor.

The key cognitive and affective states identified in the literature by Carroll, et al as key influencers of learning were measurable using only five sensors. Redundant sensors might be eliminated to reduce the sensor suite to two sensors (heart rate monitor and EEG) or be used to confirm state classifications of other sensors. While this is good news, it is not optimal. Graesser and D'Mello (2012) have identified larger sets of states that influence learning that included confusion, joy, sadness, shame, confidence, arousal, and awe. This larger set of states may require a significantly larger set of sensors. While future versions of GIFT will make recommendations for optimal sets of sensor suites for detecting specific states, GIFT has been designed to accommodate tailored selection of sensors to support the assessment of different cognitive and affective states as needed to support different types of training tasks (e.g., cognitive, affective, psychomotor, social, and hybrid).

Additional GIFT learner modeling experiments examined the effect task clarity and flow of interaction on arousal and engagement within a computer-based training environment. Goldberg & Brawner (2012), and Brawner & Goldberg (2012) monitored state changes with measures from an EEG, electrocardiogram (ECG), and galvanic skin response (GSR) sensors. Windowed time-segments were found to be significantly different across all EEG measure in each scenario condition. Analysis shows ECG data to display minimal variance over time and across scenarios. Significant differences were found for all GSR metrics examining effect of task clarity. The results discussed herein impact decisions on the type and application of sensors given the type of task and interaction with the CBTS.

Macro-adaptive strategies were employed by augmenting the scenario development pipeline to consider learner attributes during the scenario generation phase of initialization just prior to instruction. Macro-adaptive tailoring was based on previous performance assessments in lieu of developing scenarios to support average abilities of the learner population which can leave above-average learners bored and below average learners frustrated (Zook, et al, 2012). Micro-adaptive strategies are being addressed to support in-situ tailored training based on skill maps of the learner, and their expected and actual responses to skill events in the training.

Last, but not least, ARL conducted a pilot study (Holden, 2012; Holden & Goldberg, 2011) to evaluate how the tutor's competency and emotional support affect the learner's perceptions, self-efficacy, mood, and motivational level. The tutoring agent type (i.e., competent only, emotionally supportive only, competent and supportive, or neither competent nor supportive) had a very large effect on learners' perceived credibility of the agent. Learners exposed to emotionally supportive tutoring conditions reported significantly higher self-efficacy, and this may be important in training tasks where self-efficacy is a key factor in learning (e.g., where learners are domain novices). Competent tutors had a large effect on trust for the overall learning environment. This is important in that the poor instructional decisions by the tutoring agent(s) may adversely affect the learner's perceptions of the learning environment even when the instructional content (e.g., media) is credible.

Multi-Disciplinary Usability

Development time and cost are among the reasons that CBTS are not more prevalent as tools in the military training domain. Another barrier is the vast amount of expertise needed to create a CBTS and decide what should be in it. It takes multi-disciplinary teams to author CBTS, and multi-disciplinary user interfaces are needed to support the planning, development, delivery, assessment, and use of CBTS. A multi-disciplinary user-centric framework shown in Figure 4 is part of the GIFT ontology and provides for interfaces/views compatible with the knowledge base and skillset of each user allowing for a user base broader than the computer science field. The notion is to apply authoring tools to ease the cost and effort associated with CBTS development.

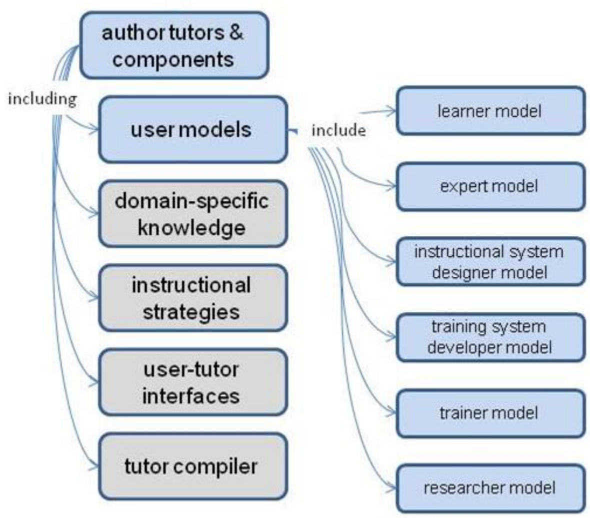

Figure 4: Multi-Disciplinary User Modeling in GIFT

Models must be developed to support each user domain. Information access, standards, look-and-feel, and even jargon should be tailored to support a variety of user disciplines. For example, GIFT is structured to aid in the development of expert (or ideal learner) models in a particular domain. The graphical user interface (GUI) to support them must provide representative domain tasks, standards, and conditions as well as automatically capture essential behaviors (e.g., decisions and actions), prompt for thoughts, and help the expert organize their knowledge.

For instructional system designers (ISDs), best design practices and instructional design models should be built into their interface or available on request. Training developers and managers would benefit from tools that can help evaluate best value to determine tradeoffs for component and system costs. The ability to organize content objects, media and other development-related data would also be a plus. Trainers/instructors also need the capability to organize information (e.g., self-development courses and lessons along with student performance data) so part of the trainer GUI design includes a learning management system interface. For researchers, GIFT can be used to analyze the comparative effect size of instructional tools and methods, and to conduct experiments to assess new technologies.

Traceability

As previously discussed, methods for assessing the learner's state are critical in selecting appropriate interventions by the CBTS and, thereby, affect learning. For example, Graesser and D'Mello (2012) determined learning gains had a significant positive correlation with confusion ($r = .33$) and flow/engagement ($r = .29$), but a negative correlation with boredom ($r = -.39$). Findings like these influence production rules integrated within GIFT and compiled into new CBTS. A mechanism to validate these production rules and then track their effect in GIFT-developed CBTS is an important capability to evolving tutoring best practices.

ARL is evolving an ontology for GIFT to provide users with traceability from empirical research to pedagogical methods adopted. Instructional strategies are in large part based on the study of expert human tutors (Lepper and Woolverton, 2002) and experiments involving learners and CBTS (Person, Kreuz, Zwaan, and Graesser, 1995). Organizing generalized strategies found to be effective regardless of the domain will assist authors in linking concepts and objectives with proven strategies for optimizing learning outcomes.

Extensibility

A CBTS has many moving parts, possibly including sensors to assess a learner, a learner model, methods for selection of instructional strategy, and implementation of instructional strategy. At any stage of GIFT development, functional modules exist as placeholders waiting to be replaced by more efficient or effective modules. Since GIFT is set up to be an interoperable service-oriented architecture, more functional modules can replace older ones assuming they support the same interactions. The same paradigm exists for sensors and services.

Research-driven changes allow for a ripple effect across the CBTS messaging pipeline. If a new sensor is available that can affordably and accurately assess learner state, this sensor can be integrated and more accurate state information can be made available to the strategy recommendation engine in the pedagogical module. Overall, GIFT could also be extended to include other messages to support future services. In this manner, one can continue to employ the 'best-available' and extend functionality as it is developed.

FUTURE RESEARCH AND DEVELOPMENT

GIFT 1.0 was released in May 2012 and is available to the Department of Defense (DoD) and DoD contractors via GIFTtutoring.org. GIFT 1.0 is based on standard messaging (JSON) sets as part of a service-oriented architecture to support comparative evaluations, but presently lacks the team models needed to support small unit tutoring. Additional research is needed to assess the effect size of team cognitive, affective, competency, trust, and communication models in support of a team performance model that assesses current performance and analyzes future team performance trends.

To aid the assessment of cognitive and affective states, two commercial sensors have been integrated in this initial version. Interfaces to the GIFT sensor module bus have been developed to integrate the Emotiv electro-encephalograph (EEG) to support cognitive engagement and workload assessment, and the Affectiva Q Sensor, a wrist-worn electro-dermal activity (EDA), skin temperature, and acceleration sensor to support assessment of arousal. Additional sensors are planned for integration based on recent experiments which assessed the ability to accurately classify cognitive and affective states (Carroll, et al, 2011).

Research is needed to develop best practices to assign individual and groups of sensors to learning domains. For example, tasks that involve primarily cognitive learning may be readily trained in a stationary mode (e.g., desktop or laptop computer) which is compatible with EEG sensors which are prone to interference/noise from learner movement. EEGs and other sensors may or may not be supported in more kinetic modes where learners move frequently and abruptly (e.g., psychomotor tasks).

To feed the CBTS assessment process, we will need models, algorithms, and methods to evaluate training effectiveness and efficiency. Authoring tools and standards will be a critical element in feeding the CBTS assessment process. Tools are needed to support the rapid generation of CBTS technologies that include expert models, learner model variants, instructional methods, and performance assessment tools.

A set of authoring goals for GIFT has been adapted from Murray (1999):

- Decrease the effort (time, cost, and/or other resources) for authoring and assessing CBTS;
- Decrease the skill threshold by tailoring tools for specific disciplines to author, assess and employ CBTS;
- Provide tools to aid the designer/author/trainer /researcher organize their knowledge;
- Support (i.e. structure, recommend, or enforce) good design principles (in pedagogy, user interface, etc.);
- Enable rapid prototyping of CBTS to allow for rapid design/evaluation cycles of prototype capabilities.

Additional research is needed to support the CBTS authoring processes. Automating the authoring processes for these technologies will provide additional fodder for the GIFT spiral development process and over time improve the generalizability, efficiency, reuse, and learning effectiveness of CBTS produced by GIFT.

CONCLUSIONS

This paper reviewed key areas to support authoring and assessment for adaptive CBTS. Today, few generalizable CBTS authoring tools exist and no assessment standards have been developed to promote reuse among CBTS. CBTS are currently not on a par with the expert human tutors that Bloom (1984) determined to have a learning effect of two standard deviations over traditional classroom training. To move toward and beyond a two sigma learning gain, CBTS will require an assessment framework to empirically evaluate tutoring technologies and evolve "best in class" tutoring components, tools, and methods. The authors have put forward GIFT as a potential CBTS assessment tool based on its modular design.

Current CBTS are not designed to support small unit training/tutoring experiences. Additional research is needed to develop and/or assess team tutoring models and constructs. Finally, CBTS do not translate well beyond desktop learning (e.g., mobile learning, mixed reality or live training). Significant work lies ahead to develop methods to support adaptable computer-based tutoring in these training domains.

REFERENCES

Anderson, L. W., and Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives: Complete edition, New York : Longman.

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(6), p. 6.

Branson, R. K., Rayner, G. T., Cox, J. L., Furman, J. P., King, F. J., and Hannum, W. H. (1975). Interservice procedures for instructional systems development (5 vols.) *TRADOC (Pam 350-30)*. Ft. Monroe, VA: U.S. Army Training and Doctrine Command, August 1975.

Brawner, K. and Gonzalez A. (2011). Realtime Clustering of Unlabeled Sensory Data for User State Assessment. International Defense & Homeland Security Simulation Workshop in Proceedings of the I3M Conference. Rome, Italy, September 2011.

Brawner, K. and Goldberg, B. (2012). Realtime Monitoring of ECG and GSR Signals during Computer-Based Training. In Proceedings of the Intelligent Tutoring Systems (ITS) 2012 Conference, Chania, Crete.

Carroll, M. Kokini, C., Champney, R., Fuchs, S., Sottilare, R. and Goldberg, B. (2011). Assessing Trainee States in Intelligent Tutoring Systems: Using Low Cost Sensors for Affective and Cognitive State Modeling. *Interservice/Industry Training Systems & Education Conference*, Orlando, Florida, December 2011.

Cohen, P.A., Kulik, J.A., and Kulik, C.L.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237-248.

Gagné, R. M. (1985). The conditions of learning (4th ed.). New York: Holt, Rinehart & Winston.
31.

Gagné, R. M., & Driscoll, M. P. (1988). Essentials of learning for instruction. Englewood Cliffs, NJ: Prentice-Hall.

Goldberg, B. & Brawner, K. (2012, Accepted). Efficacy of Measuring Engagement during Computer-Based Training with Low-Cost Electroencephalogram (EEG) Sensor Outputs. In Proceedings of Human Factors and Ergonomics Society Annual Meeting (HFES2012). Boston, MA.

Goleman, D., (1995). Emotional Intelligence. Bantam Books: New York.

Graesser, A.C. and D'Mello, S. (2012). Emotions During the Learning of Difficult Material. In: Brian H. Ross, Editor(s), Psychology of Learning and Motivation, Academic Press, Volume 57, Chapter 5, Pages 183-225, ISSN 0079-7421, ISBN 978012-3942937, 10.1016/B978-0-12-394293-7.00005-4.

Hanks, S., Pollack, M.E. and Cohen, P.R. (1993). Benchmarks, Test Beds, Controlled Experimentation, and the Design of Agent Architectures. *AI Magazine* Volume 14 Number 4.

Holden, H. (2012). Understanding the Interaction Behavior of Pedagogical Agents' Emotional Support and Competency on Learner Outcomes and Agent Perceptions. In Proceedings of the Intelligent Tutoring Systems (ITS) 2012 Conference, Chania, Crete.

Holden, H. and Goldberg, B. (2011). Student Expectations and Tutor Acceptance: The Impact on Student Affect (Mood) and Future Usage Intentions. International Defense & Homeland Security Simulation Workshop in Proceedings of the I3M Conference. Rome, Italy, September 2011.

Kemp, J.E., Morrison, C.R. and Ross, S.M. (1994). Designing effective instruction. New York: Merrill.

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

Krathwohl, D.R., Bloom, B.S., and Masia, B.B. (1964). Taxonomy of Educational Objectives: Handbook II: Affective Domain. New York: David McKay Co.

Lepper, M. and Woolverton, M. (2002). The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. In *J. Aronson (Ed) Improving academic achievement: impact of psychological factors on education* (pp. 135-158). New York: Academic Press.

Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (1988). SHERLOCK: A coached practice environment for an electronics troubleshooting job.

In Computer assisted instruction and intelligent tutoring systems. Hillsdale, NJ: Lawrence Erlbaum.

Merrill, M. D., Drake, L., Lacy, M. J., Pratt, J., & ID2_Research_Group. (1996). Reclaiming instructional design. Educational Technology, 36(5), 5-7.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. International Journal of Artificial Intelligence in Education, 10(1):98–129.

Ogan, A., Finkelstein, S., Walker, E., Carlson, R., and J. Cassell. (2012). Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring. Proceedings of the 11th International Conference on Intelligent Tutoring Systems. Chania, Crete, Greece. pp. 11-21

Park, O.-c., & Lee, J. (2004). Adaptive instructional systems. In D. H. Jonassen (Ed.), Handbook of Research for Education Communications and Technology (pp. 651-684). Mahwah, NJ: Erlbaum.

Person, N. K., Kreuz, R. J., Zwaan, R. A., and Graesser, A. C. (1995). Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. Cognition and Instruction, 13(2), 161–188.

Reigeluth, C.M. (1999). The elaboration theory: Guidance for scope and sequence decisions. In C.M. Reigeluth (Ed.), Instructional-Design Theories and Models: A New Paradigm of Instructional Theory. (Volume II). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Robinson, D., Schofield, J., and K. Steers-Wentzell. (2005). Peer and Cross-Age Tutoring in Math: Outcomes and Their Design Implications. *Educational Psychology Review*, 17(4), pp. 327-362.

Rodríguez, J. I., Plax, T. G., & Kearney, P. (1996). Clarifying the relationship between teacher nonverbal immediacy and student cognitive learning: Affective learning as the central causal mediator. *Communication Education, 45*(4), 293-305.

Rohrbeck, C.A., Ginsburg-Block, M.D., Fantuzzo, J.W., Miller, T.R.: Peer-assisted learning interventions with elementary school students: a meta-analytic review. Journal of Educational Society 95(2), 240–257 (2003)

Sharpley, A., Irvine, J., and Sharpley, C.: An examination of the effectiveness of a cross-age tutoring program in mathematics for elementary school children. American Educational Research Journal 20(1), 101–111 (1983).

Simpson, E. (1972). *The classification of educational objectives in the psychomotor domain:* The psychomotor domain. Vol. 3. Washington, DC: Gryphon House.

Soller, A. (2001). Supporting social interaction in an intelligent collaborative learning system. *International Journal of Artificial Intelligence in Education*, 12(1), 40-62.

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

Sottilare, R., Holden, H., Brawner, K. and Goldberg, B. (2011). Challenges and Emerging Concepts in the Development of Adaptive, Computer-based Tutoring Systems for Team Training. *Interservice/Industry Training Systems & Education Conference*, Orlando, Florida, December 2011.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15(3), 147–204.

Webb, N.: Peer interaction and learning in small groups. International Journal of Educational Research 13(1), 21–39 (1989).

Woolf, B.P. (2011). Intelligent Tutors: Past, Present and Future. Keynote address at the *Advanced Distributed Learning ImplementationFest*, August 2011, Orlando, Florida.

Zook, A., Riedl, M., Holden, H., Sottilare, R., and Brawner, K. (2012). Automated Scenario Generation: Toward Tailored and Optimized Military Training in Virtual Environments. Foundations of Digital Games Conference (FDG 2012). Raleigh, NC.