Automated Assessment of Team Performance Using Multimodal Bayesian Learning Analytics

Caleb Vatral, Gautam Biswas, Naveeduddin Mohammed
Vanderbilt University
Nashville, TN
caleb.m.vatral@vanderbilt.edu,
gautam.biswas@vanderbilt.edu,
naveeduddin.mohammed@vanderbilt.edu

Benjamin S. Goldberg
US Army CCDC Soldier Center
Orlando, FL
benjamin.s.goldberg.civ@mail.mil

ABSTRACT

This paper presents a case study of teams of soldiers training on dismounted battle drills in a mixed-reality training environment. Mixed-reality simulation-based training environments along with multimodal sensing devices have made it much easier to collect and analyze participant interaction and behavior data for evaluation and feedback. Advanced AI and machine learning algorithms have further enhanced the ability to create robust multi-dimensional individual and team performance models. The performance metrics computed within single training instances, can be extended to cover a full course of training scenarios. This provides valuable feedback to trainees and their instructors on their skill levels across cognitive, metacognitive, affective, and psychomotor skill dimensions. However, developing objective data-driven performance metrics comes with a set of challenges that includes data collection and aggregation, pre-processing and alignment, data fusion, and the use of multimodal learning analytics (MMLA) algorithms to compute individual and team performance. We develop a generalized multilevel modeling framework for the training domain and use machine learning algorithms to analyze the collected training data that spans video, speech, and simulation logs. We model teams of soldiers through multiple training scenarios and show their progression over time on both operationalized domain-specific performance metrics, as well as higher-level cognitive and metacognitive processes. We conclude with a discussion of how results from our analysis framework can be used to provide formative feedback to trainees and suggestions for future training needs, as well as data-driven evidence to be used as part of a longer-term summative assessment system.

ABOUT THE AUTHORS

Caleb Vatral is a PhD student and research assistant at Vanderbilt University in the Department of Computer Science with a focus in intelligent systems. Working at the Institute for Software Integrated Systems, his research focuses on combining theoretical foundations in distributed cognition with multimodal data-driven approaches to support cognitive modeling and system design in human-centered simulation and simulation-based training. Prior to attending Vanderbilt, he received the B.S. degree in computer science and mathematics from Eastern Nazarene College.

Dr. Gautam Biswas is a Cornelius Vanderbilt Professor of Engineering and Professor of Computer Science and Computer Engineering at Vanderbilt University. He conducts research in Intelligent Systems with primary interests in monitoring, control, and fault adaptivity of complex cyber physical systems, as well as developing intelligent open-ended learning environments that adapt to students' learning performance and behaviors. He has developed innovative multimodal analytics for studying students' learning behaviors in a variety of simulation and augmented reality-based training environments. He has over 600 refereed publications, and his research is supported by funding from the Army, NASA, and NSF.

Naveeduddin Mohammed is a Senior Research Engineer with the Institute for Software Integrated Systems at Vanderbilt University. Naveed received the M.S. degree in Computer and Information Sciences from University of Colorado. He is a full stack developer, and his work focuses on designing, developing, and maintaining frameworks for open-ended computer-based learning environments and metacognitive tutors.

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 the Synthetic Training Environment. 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.

Automated Assessment of Team Performance Using Multimodal Bayesian Learning Analytics

Caleb Vatral, Gautam Biswas, Naveeduddin Mohammed
Vanderbilt University
Nashville, TN
caleb.m.vatral@vanderbilt.edu,
gautam.biswas@vanderbilt.edu,
naveeduddin.mohammed@vanderbilt.edu

Benjamin S. Goldberg
US Army CCDC Soldier Center
Orlando, FL
benjamin.s.goldberg.civ@mail.mil

INTRODUCTION

Effective lifelong training is a critical component of success for complex workplace tasks. Training programs and technologies are designed for learners to practice a variety of cognitive, psychomotor, affective, and metacognitive skills in safe environments. One of the critical components for effective learning outcomes is proper assessment (Pierce, 2002; Tosuncuoglu, 2018). However, traditional methods for assessment have relied on observations by domain experts. This has its limitations, including the high cost of having experts present across multiple training instances, the difficulties experts may have of remembering nuanced details of trainees' activities, and the potential biases and judgmental differences introduced by each individual assessor.

Motivated by these issues, in this paper we develop a framework for automated team performance assessment based on analysis of multimodal trainee data. Our methods use Cognitive Task Analysis (CTA) and Bayesian inference methods to generate evaluations that range from domain-specific to high-level domain-general concepts. Automating evaluations supports training analyses without expert intervention, thus increasing the objectivity and robustness of the assessments, and the feedback that can be generated from the assessments. In addition, our automated schemes are data-driven and evidence-centered, making them repeatable and decreasing the potential for biases.

While our analysis framework is extensible across all LVC[1] training, our initial development of the framework focuses on mixed-reality simulation environments (MRSEs). MRSEs offer unique advantages to virtual training. By using a mix of physical and digital elements, they keep the benefits of digital training, including safety and repeatability, while also maximizing psychomotor and cognitive experiences as trainees move around and physically interact in the space. However, this mix of physical and digital requires sophisticated analysis techniques that can link events across the physical and digital spaces of the environment. Therefore, multimodal data collection and analysis becomes a prerequisite for automated evaluation of trainee performance. An added advantage of MRSEs is that it makes our methods transferrable to digital (i.e., *constructive* and other *virtual* training) and physical (i.e., *live* training) environments.

We demonstrate our analysis framework using a case study of fire team training in an MRSE. By analyzing the environment's multimodal data, we track performance of soldiers over time and show how this progression provides insights into the personalization of current and future training of skills and the optimal length of training sessions to achieve effective learning gains. Along with the design, description, and demonstration of the automated analysis framework, we investigate the following research questions:

1. How do we develop hierarchical teamwork model structures that provide a mapping from observable and measurable behaviors to higher-level teamwork skills in relation to the overall training task?
2. How can the automated data-driven assessments of team performance be utilized by trainees and instructors to focus the training experience on specific trainee needs and improve overall learning outcomes?

[1] The US Army, classifies training systems into the Live (real people training on real systems and equipment), Virtual (real people training on simulated systems and equipment), and Constructive (simulated people operating simulated systems and equipment), i.e., the LVC taxonomy (Modeling and Simulation Glossary, 2011).

BACKGROUND

Mixed-Reality Simulation-Based Training

The application of simulation and MRSEs is not a new concept to the Armed Forces. eXtended Reality (XR) solutions merge physical task execution with interactive synthetic resources to create sufficient fidelity and immersion to elicit realistic behavioral responses. They have been applied successfully in early-exposure and refresher training across complex and dangerous task domains for combinations of cognitive, psychomotor, and affective skills. These approaches also offer unique training opportunities by rapid exposure to multiple scenarios under safe, controlled conditions. These mixed reality methods have been successfully implemented across multiple Programs of Instruction, with well-documented impacts in marksmanship (Debeltz, 2017) and medical (Barrie, et al., 2019) training.

MRSEs have matured significantly over the last decade across wireless computing technologies, wearable and behavioral sensors, game engine mechanics, motion and weapon tracking fidelity, and visualization tools to support playback and After-Action Review (AAR). The Army's Synthetic Training Environment (STE) modernization program (Goldberg et al., 2021; *Synthetic Training Environment*, n.d.) aims to take advantage of these state-of-the-art XR capabilities to deliver effective collective training at lower echelons that focus on deliberate practice of complex tasks (Ericsson, 2009). A requirement within STE is to establish training management tools that leverage multimodal data produced across XR environments. This will enable automated assessments and evidence-based intelligent tutoring functions to track performance and proficiency over time and optimize learning outcomes. These environments can combine multiple streams of data (e.g., video, audio, simulation, game-state interaction, physiological, and eye tracking data along with other behavioral signals) to develop a more comprehensive and objective understanding of the actions and decisions of a trainee or team during execution of a task. To facilitate extensible solutions leveraging these data types, research is required to establish a data architecture and a set of workflows to manage real-time capture of multimodal data, and analyses of this data using AI and Machine Learning algorithms to support data-driven performance assessments.

Cognitive Task Analysis

In complex training domains, CTA methods are commonly used to decompose complex tasks and behaviors into their component sub-tasks (Clark & Estes, 1996; Zachary, et al., 2000). CTA models are hierarchical; tasks and concepts at the highest-levels represent general cognitive processes, and each deepening level contains more domain-specific actions and behaviors. The models are constructed by iterative refinement; each task is broken down into its component sub-tasks at the next level of the hierarchy. Sub-tasks are then further deconstructed into more fine-grained subtasks, until the leaves capture observable actions and behaviors. Our CTA models constructed by task decomposition include a thorough review of army doctrine, structured interviews with domain-experts and instructors, and observations of trainees performing their tasks.

Our prior work on learner modeling in simulation-based training domains exploited the hierarchical structure of the CTA model to generate inferences about domain-general cognitive processes using observable, situation-specific low-level data (e.g., Kinnebrew, et al, 2017; Zhang, et al., 2021). Biswas et al. (2020) used a CTA approach to analyze trainee performance in a counterterrorism simulation called *UrbanSim*. By logging trainee actions in the simulation environment, we monitored their performance across multiple levels of abstraction using quantitative measures to capture their cognitive and metacognitive processes. In Vatral et al. (2021), we applied CTA to generate a hierarchical model and associated quantitative metrics for the *Enter and Clear a Room (ECR)* dismounted battle drill. Continuing this work, Vatral et al. (2022a) generalized the CTA methods to examine teamwork behaviors in addition to individual performance. We created the Hierarchical Affect, Behavior and Cognition (H-ABC) model of teamwork that linked high-level teamwork concepts to domain-specific performance metrics using data collected from the training simulation. By propagating these performance metrics in the hierarchical model, we generated automated evaluations of teamwork in the ECR domain. Furthermore, using a case-study approach, we showed that our automated evaluations of teamwork matched instructor evaluation and feedback. In this paper, we extend the H-ABC model to analyze team performance in the ECR domain.

Multimodal Analytics

As discussed in the previous section, we use multimodal sensors, such as cameras, microphones, and chest harnesses to collect soldier activity data, which is then processed using the CTA model to derive a set of performance metrics. Analysis of multimodal data is a growing field that provides the basis for generating holistic inferences of trainee actions, behaviors, and affective states (Blikstein & Worsley, 2016). Multimodal analysis is even more important in mixed-reality team training environments. For example, with unimodal analysis, we might collect audio transcripts to look for patterns of communication among team members. However, even for simple commands, such as "Go check that," we cannot infer its meaning from the transcript alone. An additional data modality, such as video may reveal a pointing gesture along with the command that allows us to disambiguate what "that" means in this context. Multimodal data facilitates a more complete analyses of trainee psychomotor, cognitive, and affective states (Ochoa et al., 2017).

In our work, we adopt a *late fusion* approach to multimodal analysis (Sleeman et al., 2021). Individual performance metrics are calculated using single data modalities, and then multiple metrics are fused together using Bayesian network models to generate the performance assessments at higher levels of the CTA hierarchy. Bayes nets are graphical probability models that allow us to infer unknown variable values using combinations of known variable values and conditional probability distributions that establish the relations between variables in the graph (Ben-Gal, 2008). In our application, we use the observable data and the performance metrics derived from that data to infer performance values for the unobservable higher-level concepts. Our Bayes net follows the structure of the CTA model for the domain. Each performance construct can be in one of three possible states: *below*, *at*, and *above expectation*. This three-state learner model matches the design of the Generalized Intelligent Framework for Tutoring (GIFT) (Goldberg, et al., 2021), and is analogous to the apprenticeship model often seen in Army training doctrine and psychology literature on expertise development (e.g., *Novice*, *Journeyman*, *Master*; or *Crawl*, *Walk*, *Run*) (Cassella, 2010; Klein & Hoffman, 1992; Sottilare, et al., 2017). We generated the probability distributions for the Bayes net by consulting domain experts and reviewing training doctrine. These distributions can be further tuned using trainee data. We reserve this for future work, as it requires more data than is available from our case study described in the next section.

CASE STUDY

To drive the development and validation of the analysis methods presented in this paper, we use data collected from a case study of two infantry fire teams that participated in a study at Fort Campbell over the course of two days. The teams trained on the ECR dismounted battle drill using the Squad Advanced Marksmanship Trainer (SAM-T). In the drill, the fire team enters a room with the goal of neutralizing all enemy combatants. Each room may contain any combination of enemy combatants, civilian non-combatants, physical obstacles, and unique weapons. From a doctrinal view, soldiers train to rapidly enter the room one after another, following paths of least resistance along the walls and neutralizing enemy combatants in their sectors of fire. Once all combatants are neutralized and all civilians are secured, team members, directed by the team leader successively search each entity to remove weapons. Once every entity is searched and neutralized, the team leader gives the all-clear signal, and the team exits the room, vocalizing their exits to ensure that no fratricide occurs.

During each scenario, soldiers must also be aware of variable task conditions that dictate their behavior and modifications to the standard protocols. For example, situation changes may include items such as new combatants entering the room, civilians being noncompliant or becoming aggressive, or the presence of explosive devices in the room. The ECR scenario in SAM-T uses a U-shaped arena with three screens, on which a Virtual Battle Simulator 3 (VBS3)-generated scenario that evolves in the room is projected. The soldiers move around in the space created by the U-shaped arena and fire their weapons at enemy on-screen entities. The SAM-T system logs a variety of data related to both soldier and entity actions in the VBS3 scenario. Logged events include soldier weapon firing, Zephyr biometric harness data, VBS3 entity weapon firing, and VBS3 entity positions. In addition, we augmented the SAM-T environment to include two video cameras that captured soldier movements and actions from two distinct angles.

We used the GIFT data collector on-site at Fort Campbell for live, synchronized data collection between SAM-T, VBS3, and our external cameras. The data recorded in log files can be played back through GIFT after the study. All our assessments were conducted offline after the study, but teams still received normal assessment feedback from the instructor during the training. We recoded the instructor feedback and the discussions among team members for subsequent analysis. Each drill took between 30 seconds and 2 minutes to complete, and each team completed between 20 and 30 drills. Intermittent sensor failure during the study affected some of the weapon firing, Zephyr bio harness, and video camera data collection, thus resulting in some data loss. In total, we used complete data from 41 drills for the analysis in this paper.

THEORETICAL FRAMEWORK

Figure 1 shows our theoretical framework for automated assessment of team performance. Overall, it follows a generic input-process-output (IPO) structure, i.e., the model describes task execution as three modular components: (1) *Input* that represent the basic structure of the team, including individual personalities and knowledge, skills, and attitudes (KSAs), as well as more organizational concerns, such as the resources available to the team and the task context in which the team operates; (2) *Processes* that represent the actions taken by the team during task execution to accomplish their shared goals; and (3) *Outputs* that represent the outcomes of the team operation including actualized performance with respect to both individual and team goals.

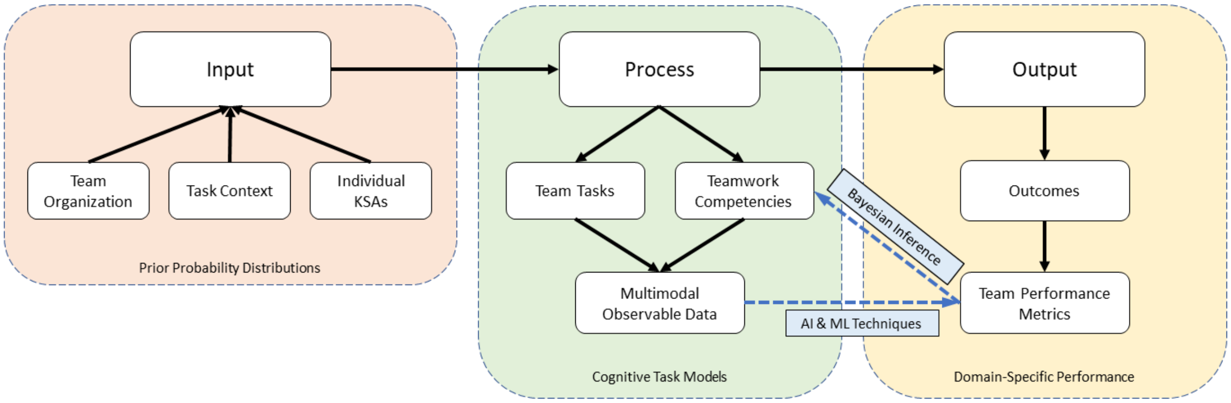

Figure 1. Overall framework to compute team performance from observable data in terms of the IPO model. Solid arrows represent causal relationships and dashed arrows represent computational processes

The Three IPO Components

In general, outputs represent a wide variety of outcome concepts including how well the team achieved their shared goals, whether the team followed proper protocols and best practices, the viability of the team for future training and missions, and perceptions of competency and efficacy. To describe these general outcome concepts, we developed a set of domain-specific performance metrics to score the outcomes as numeric measures. For example, army doctrine for ECR, suggests that the team should move along the walls of the room to minimize the blind spots at their backs.

Table 1. The five domain-specific performance metrics which are automatically calculated from the ECR case-study data and used in the cognitive task model.

Metric Name	Description	Calculation
Points of Domination (PODs)	How well soldiers reach and maintain their PODs?	Normalized minimum Euclidean distance between soldiers and their PODs
Move Along Wall	How well do soldiers keep along the walls of the room while entering?	Percentage of video frames where soldiers are within a distance threshold of the wall
Entrance Vectors	Do the soldiers enter the room and move in the opposite direction of the previous soldier?	Percentage of soldiers for whom the angle of their entrance vector is opposite of the previous
Total Entry Time	How quickly does the team enter the room once commenced?	Normalized difference between team's entry time compared to the optimal time threshold
Entrance Hesitation	How quickly does each soldier enter the room after the previous soldier?	Normalized difference in entry time between two successive soldiers compared to an optimal time threshold

Given this doctrinal best practice, we designed a performance metric to measure movement along the wall by comparing the average distance the trainee maintained from the wall to an optimal threshold value. Regardless of the specifics of each training domain, the metrics need to be directly computable from the collected data. For example, to compute the move along wall metric, we can use computer vision techniques applied to collected video. In our case, we use a number of machine learning and AI techniques to compute these domain-specific performance metrics, represented as the blue dashed arrow connecting *observable data* and *team performance metrics* in Figure 1. For our case study, we defined five domain-specific performance metrics that were calculated automatically (see Table 1).

The process component represents the behaviors and actions taken by the team during task execution to accomplish their individual and shared goals. These actions and goals vary widely depending on the domain being analyzed, but there are also shared commonalities in high-level cognitive concepts and tasks that extend across domains. To facilitate both the commonalities and differences between analyzed domains, we model the process component using CTA (Zachary, et al., 2000). In this work, we adopt the H-ABC model of teamwork as the highest levels of the task analysis. In Vatral et al. (2022a), we extend the H-ABC model by one additional layer containing the domain-specific performance metrics described in Table 1. We link each performance metric to concepts in Level 3 of the model by analyzing the teamwork components that contribute to performance of the metric (see Figure 2). For example, a metric that measures the hesitation of soldiers as they enter the room is linked to the *Backup Behavior* and *Task Comprehension* concepts in Level 3 of the H-ABC model. With the five team performance metrics that we used for this case study, not all the H-ABC model concepts are covered by links to the primary performance metrics. For example, affective components of the H-ABC model such as *conflict resolution* and *mutual trust* do not have any associated performance metrics that can be computed from the data collected in the case study. In future work, we will focus on collecting more data modalities to allow automated computation of additional teamwork competencies.

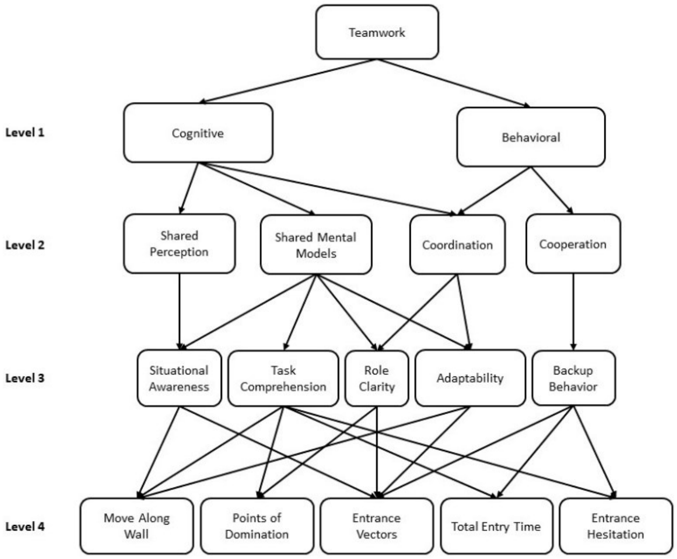

Figure 2. The cognitive task model used for this study

In our previous work, we propagated the performance metrics up to higher levels of the CTA hierarchy by simply averaging the performance linked to each higher-level concept (Vatral et al., 2022a). While this simple rollup method showed promising results, it does not fully capture the complex interdependencies between teamwork concepts. In this work, we enhance the H-ABC model using Bayesian inference to propagate the performance metrics up to higher levels. Using the probabilistic graphical structure, we can derive the conditional distributions that capture complex interdependencies between teamwork concepts. For each concept, we model the trainee state using the three-state learner competency model*:* *below*, *at*, and *above expectation* as discussed above. In addition, we define a transition model, which represents the probability of transitioning between each of these three states after a training event, and a conditional probability model, which propagates the probability from lower-level concept values to the states of their higher-level parents. By defining these two probability models in the Bayes net, we can track the progression of the performance metrics over time and show how trainees transition between each of the three learner states over the course of the training program. Figure 1 illustrates this probabilistic propagation and inference of higher-level concepts states by the blue arrow connecting *team performance metrics* and *team competencies*.

Finally, the input component represents the static inputs to the system at the beginning of a state of training episodes. This includes concepts such as the basic structure of the team, individual personalities, prior KSAs, the resources available to the team, and the task context in which the team operates. We represent this input component by prior probability distributions in the Bayes net model that account for varying skill levels and prior experience. For example, new soldiers will most likely be in the below-expectation (novice) state for most competencies, with perhaps a small probability that they are in the at-expectation (journeyman) or above-expectation (master) state. In contrast, an experienced infantry team that has previously trained together will most likely be in the at or above expectation state for their competencies, with a small chance of still being in the below-expectation (novice) state. To drive development, the persistent representation of performance is guided by standards and best practices implemented in the Army's STE Experiential Learning for Readiness (STEEL-R) project (Goldberg et al., 2021).

Complete Computational Framework

Following our discussion of the IPO components, we present the complete computational workflow for automated assessment of trainee performance. At the start of the analysis, we define the prior probability distributions of each concept in the H-ABC model using persistent performance records. As the team completes training exercises, observable data is analyzed using machine learning and AI methods to compute the domain-specific performance metrics. These metrics are fed into the Bayes net model to compute the updated competency-state and probabilities for each teamwork concept. The performance metrics become the evidence variables for the Bayesian inference of teamwork concepts. This process then repeats itself, this time with the computed competency-state probabilities from the last training instance representing the prior probability input for the next training instance. As the team continues to complete training scenarios, the model continues to be updated with the new evidence and the confidence in the assessment of team performance increases. We can then track the progress of the trainees over time across levels of abstraction in the H-ABC model. Next, we demonstrate the application of this computational process to the ECR case-study and illustrate the performance progression of the two teams over the course of one full day of training.

RESULTS

We breakdown the analysis and results into two sections. First, we describe ECR domain-specific performance metrics used as the lowest level of the task model for our case study and the Bayesian network model used to propagate these metrics up multiple levels of the ABC hierarchy. Next, we show the results of the performance metrics, and how performance of the fire teams progresses over time at multiple levels of abstraction.

Performance Metrics Calculation and Propagation

For each of the five ECR-specific performance measures described in Table 1, we calculate a metric as a continuous value on a scale from 0 to 1 (Vatral, et al. 2021; 2022a). To convert these continuous numeric metrics to the three-state learner competency model, we have applied a simple thresholding rule. We labeled scores under 0.4 as *below-expectation*, scores from 0.4 to 0.9 as *at-expectation*, and scores above 0.9 as *above-expectation*. More advanced methods, such as Bayesian knowledge tracing (e.g., Vatral et al., 2022a), may be used to adaptively convert raw metric scores to learner competency states, but this is currently reserved for future work. In practice, this simple thresholding rule generates promising results that we describe in the next subsection.

Once we have the five performance metrics calculated for a given run of the ECR scenario, we use the Bayes net model previously described to roll-up these low-level metrics to higher-levels of the H-ABC model. For this case study, we hand-designed the probability parameters for this model based on discussions with domain-experts and careful inspection of the data. We specified three probability distributions. First, the *transition model* specifies the probability of transitioning from one competency state to another (e.g., below-expectation to at-expectation) after each training instance. As trainees perform multiple training instances (as in our case-study), the probability of transitioning can go up or down depending on their current performance. Second, the *conditional model* specifies the probability of a parent concept in each competency state given the probability states of its children. For example, if the child, *Backup Behavior* is in the below-expectation state, it is very likely that *Cooperation* is also in the below-expectation state, with a small probability that it is in the at-expectation state and an even smaller probability that it is in above-expectation state given the values of all other child nodes. In cases where a child has two or more parents, the full conditional probability distribution is constructed by making the independence assumption (i.e., by multiplying the conditional probability distributions of each of its parents). For example, *Situational awareness (SA)* has two parent nodes in the H-ABC model, *shared perception (SP)* and *shared mental models (SM)*, so its full probability model is the multiplication of its individual parents' conditional models, i.e., $P(SA\ |SP, SM) \propto P(SA\ |SP)\ \times P(SA\ |\ SM)$. Finally, the *prior model* input specifies the competency state probability for each H-ABC concept at the start of this training episode, as described in the last section. For our case study, the fire teams had varying prior experience with the ECR drill, and they had not previously trained in the SAM-T environment. Since we did not have much information about their prior knowledge, we initialized the model with all teams set to the below-expectation state with 100% probability at the start of training.

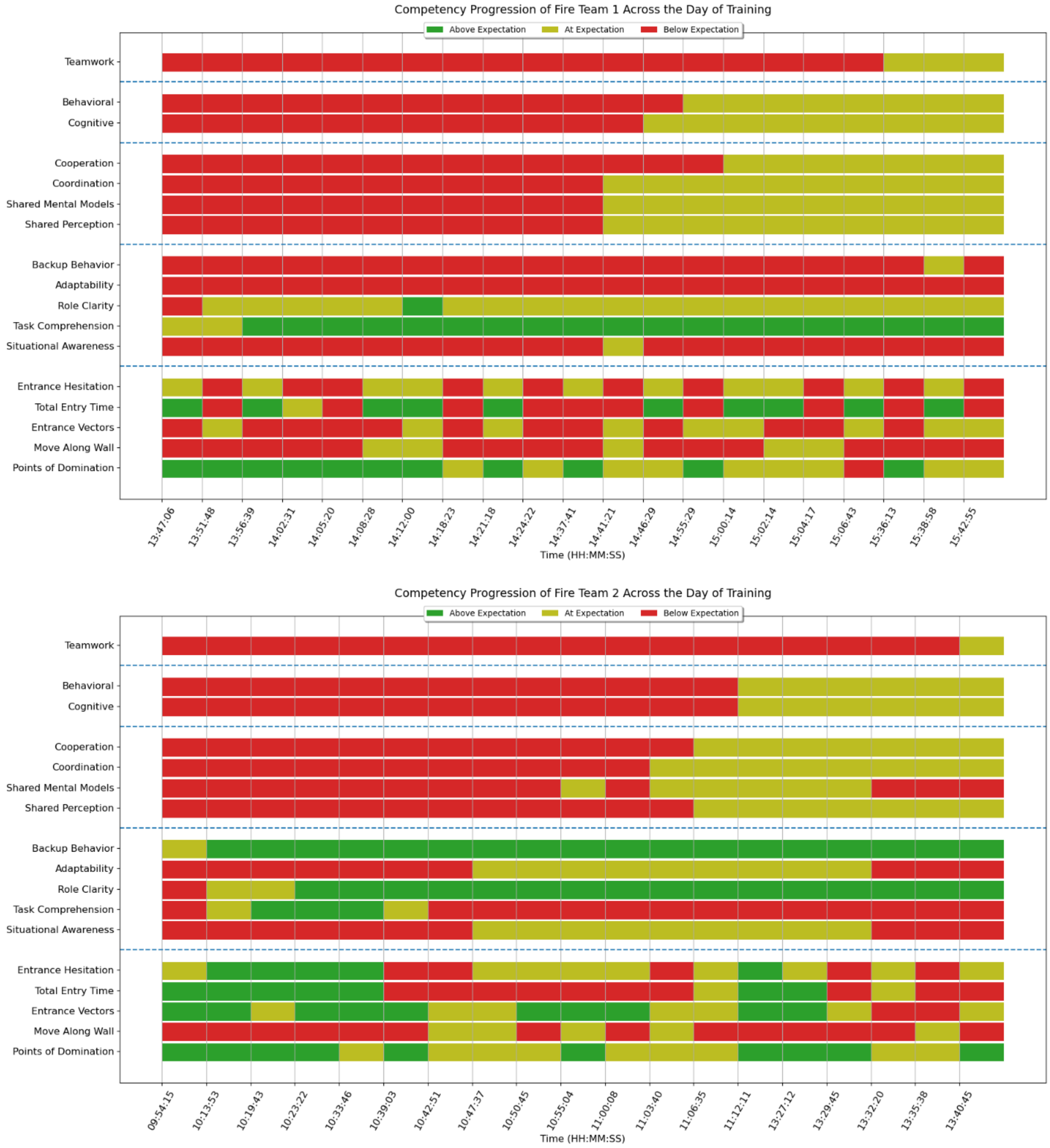

Figure 3. Performance progression across each level of abstraction in the H-ABC model over the course of the entire training day for the two fire teams who participated in the case-study.

Performance Progression

Using the domain-specific metrics and Bayes net propagation through the H-ABC model, we accumulated the performance of each fire team at multiple levels of abstraction for an entire day of training. Figure 3 shows the performance progression for the two fire teams. Three colors represent the competency states – red for below-expectation, yellow for at-expectation, and green for above-expectation. The results at each level of the H-ABC model are marked by horizontal dashed lines. From this performance progression data, three major themes emerge.

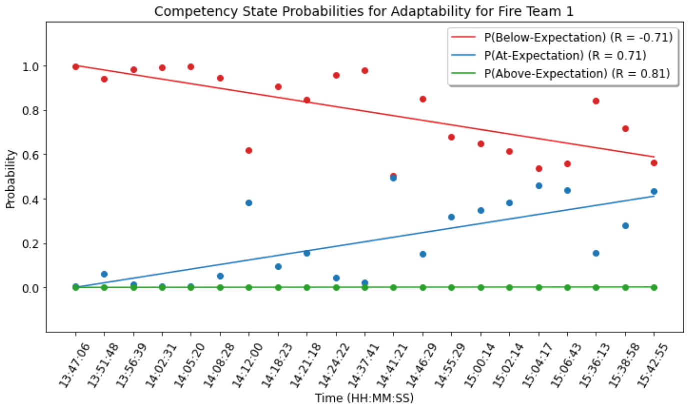

Figure 4. Progression of the competency state probabilities for the *Adaptability* concept for fire team 1

First, both teams improved as they repeated the training throughout the day. For both teams, a number of concepts increased to at- or above-expectation by three-quarters of the way through the training session. By the end of the day, there were three to four H-ABC concepts that were still below-expectation for the two teams. However, these concepts also showed improved performance through the day. For example, Figure 4 show the Bayes net probability for each competency state for the *Adaptability* concept for fire team 1. By the end of the day this concept was still below-expectation, but there was a clear upward trend in the probability of at-expectation (R = 0.71) and a clear downward trend in the probability of below-expectation (R = −0.71). This indicates that while the team did not fully grasp this concept enough to move to the next expectation level, there is evidence that their performance was improving and that with further training, the team would eventually grasp this concept at the at-expectation level.

Second, both teams mastered the lower-level concepts first, which then led to their mastering the related high-level concepts later in the day. For example, fire team 1 moved to at-expectation in *Role Clarity* and *Task Comprehension* very quickly after the training commenced, but *Shared Mental Models,* which has role clarity and task comprehension as child concepts, did not move to at-expectation until halfway through the training session. This trend is partially a result of the Bayes net model, which requires more evidence to support the higher-level concepts due to their distance from the observable performance metrics. However, this also mimics what we should expect to see from trainees. Mastering performance on one specific task (such as ECR) is much easier than mastering a domain-general high-level concept, so we should require more performance evidence from the trainees to justify that they have mastered these high-level concepts.

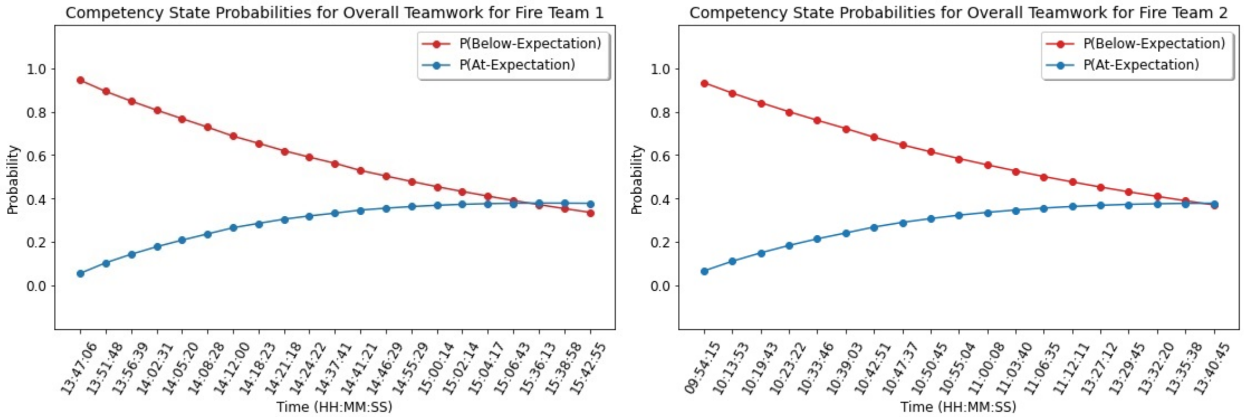

Figure 5. Competency state probabilities of below- and at-expectation for overall teamwork

Third, learner competency state transitions for the higher-level H-ABC concepts are smoother than state transitions for low-level concepts. At Level 4, the trainees exhibit a significant number of state transitions, jumping back and forth between below, at, and above expectation levels between individual scenario runs. At Level 3, the state transitions become smoother, only transitioning between states a few times through the course of the training session. At Levels 2 and higher, the state transitions are very smooth, typically only transitioning a single time during training. This behavior is sensible, since acute changes in the specific ECR scenarios are very likely to affect ECR-specific performance but are not likely to greatly affect higher-level teamwork concepts which require more evidence to transition. This illustrates one of the major advantages of hierarchical modeling. By analyzing performance across multiple levels of abstractions, we can gain insights about both domain-specific and higher-level performance concepts, which are typically hidden and difficult to observe. Because of this smoothing effect, it becomes easier to observe when gains accomplished by a specific training scenario, e.g., ECR, saturate. For example, Figure 5 show the progression of probabilities for the below- and at-expectation states for overall teamwork at the highest level of the H-ABC model. The teamwork concepts follow similar progression patterns for both fire teams, first beginning with a rapid rise in the probability of the at-expectation state, indicating improving performance, before starting level-off and reach a plateau. This plateau in high-level teamwork progression indicates diminishing returns for the training and suggests that any further training will not significantly increase performance. This is highly related to the idea of spaced practice, which suggests that people learn best with repeated, spaced practice/training sessions rather than a single long practice session (Perruchet, 1989; Smolen et al., 2016).

DISCUSSION AND FUTURE WORK

In this section, we discuss the broader implications of our automated assessment framework for team training. Based on the case-study results, we highlight three potential advantages of our framework in assessing trainee competence: training session personalization; formative feedback for After-Action Review (AAR); and support for longer-term summative assessment.

First, our assessment framework supports training session personalization. One example was introduced in the previous section and highlighted in Figure 5. By examining when the learning gains for high-level H-ABC concepts begin to plateau, we can determine when a given training session should end to minimize the diminishing returns of continued training and to maximize the benefits of spaced practice. Our inferencing scheme can be further extended using additional sensors that monitor affective states, such as fatigue and stress. However, beyond simply helping to determine the optimal length of practice, the assessment framework can help identify tasks, competencies, and specific skills that trainees may need to pay more attention to. By examining the performance progression of the team across training episodes within a single session and across multiple sessions, instructors can highlight specific skills and concepts that may require additional deliberate practice. In addition, the simulation-based environments may be modified to emphasize tasks and activities that relate to the deficient skills and practices. For example, by the end of the case-study training session, fire team 1 was still below-expectation on the *Backup Behavior, Adaptability,* and *Situational Awareness* skills while most other skills were at- or above-expectation., The instructor could use this assessment to ensure that the next training session for team 1 contained scenarios focused on those three skills specifically.

Second, our assessment framework can be used to support AAR. Specifically, by using the assessments that we generate combined with models of the specific training domains and general cognitive theory, we can generate formative feedback designed to aid the discussion during AARs and provide actionable suggestions to trainees. Our preliminary work investigating the use of our assessment framework for formative feedback has focused on generating visual feedback elements. Current prototyping leverages GIFT's Game Master User Interface (Goldberg, Hoffman & Graesser, 2020). The visual elements we are developing include charts, images, and videos, that highlight trainee and team progress, trainee actions that contributed to degraded performance, and actionable insight on behaviors trainees can change to improve their performance. For example, with the *Move Along Walls* metric in the ECR domain, our visual feedback element might show a top-down map of the scenario room with overlays of the movement paths of each soldier. From this visual map, soldiers can begin to see where they may have deviated too far from the walls of the room and to reflect on how to improve on these issues with the rest of the team. The overall idea is that the assessments generated from our framework are explainable and can be used as the basis for formative learner feedback. Further discussion about our formative feedback generation using the assessment framework is presented in Vatral et al. (2022b).

Third, our assessment framework can be integrated into a larger ecosystem for performing summative assessment of trainees over the longer-term across multiple training scenarios and environments. Our assessments are evidence-based and data-driven, meaning that assessment of learner competency is repeatable and does not rely on subjective expert judgement, enabling reliable trend analysis. Because of this, the assessments will be consistent across time and across different teams, regardless of where a team trains or who the instructor is at the time, enabling reliable trend analyses across numerous scenarios and sessions. In addition, the assessments are backed by evidence, which can be played back if there are ever concerns for the validity of a given assessment. Our assessments are hierarchical, evaluating both domain-specific performance and domain-general competence. Thus, we can use evidence from multiple sources to support a common assessment and conclusion about a team. When adapting our framework to a new domain, we simply change the lowest levels of the H-ABC model to new domain-specific performance metrics, while the upper levels that measure abstract team performance remain the same. Thus, we can use multiple specific domains to provide evidence for the assessments of common high-level concepts and behaviors.

Though the initial results presented here are promising, this study has limitations and requires future work to fully develop and validate the framework. While the case study provided a promising start and initial validation of the architecture, the study focused specifically on the ECR drill on the SAM-T system. We were careful to design our framework in such a way that it will generalize to other environments and training domains, but future work will focus on validating the system using other battle drills, as well as entirely different tasks and training domains. In addition, the use of this case-study means that the framework has been developed in *playback mode*, meaning that evaluations were not generated during live training, but rather from playback of the recorded training logs and data. This was a necessary step in development, as it allowed us to rapidly iterate on prototypes without scheduling additional studies, but future work will test our automated feedback system during live training to ensure it can be used as part of a full-scale training capability. In addition, this future live testing will allow us to collect feedback from trainees and instructors about the usability and accuracy of the assessment and associated formative feedback. This feedback from stakeholders will provide valuable insights into furthering the capabilities and utility of the system for its target audience.

Finally, future work will also focus on continued development of the system's formative feedback capabilities designed to present explainable and actionable feedback to trainees and instructors, as well as continued development of the framework as part of a larger long-term summative assessment system. For example, we have begun working with the STE Experiential Learning for Readiness (STEEL-R) project (Goldberg, et al., 2021) to integrate our H-ABC learner competency framework as a model for evaluating teamwork over the long-term. It is our goal that with continued research and development, this automated assessment system can represent a comprehensive evidence-based tool used to improve learning outcomes across a variety of team training.

SUMMARY AND CONCLUSIONS

In this paper, we presented a framework for generating automated performance assessments across a generalized formalization of competency using the multimodal data collected from mixed reality and LVC supported training environments. Our framework is grounded in cognitive task analysis and structured hierarchically, allowing low-level data and domain-specific performance metrics to generate insights about higher-level cognitive, behavioral, and affective teamwork concepts using Bayesian inference across multiple levels of the hierarchy. We presented a case study of two fire teams of soldiers training on the enter and clear a room dismounted battle drill to demonstrate analysis using our automated performance assessment framework. From the case study analysis results, we showed how our framework could be used to model performance progression of the teams over time to highlight areas where the team is performing well and areas for improvement. The overall goal is that our assessment framework could be used as part of a larger training management tool to inform decisions about current and future training needs.

ACKNOWLEDGEMENTS

This research and development work has been supported in part by funding from US Army CCDC Soldier Center Award #W912CG2220001. We would like to express gratitude to the U.S. Army's PEOSTRI, Oak Grove Technologies, and Inveris Training Solutions for access to the SAMT and technical support in executing our data collections. We would also like to thank Sae Schatz, who provided feedback during several stages of the writing process and strengthened this paper greatly. The views expressed in this paper do not necessarily represent the position or policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

Barrie, M., Socha, J. J., Mansour, L., & Patterson, E. S. (2019, September). Mixed reality in medical education: a narrative literature review. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* (Vol. 8, No. 1, pp. 28-32). Sage CA: Los Angeles, CA: SAGE Publications.

Ben-Gal, I. (2008). Bayesian networks. In F. Ruggeri, R. S. Kenett, & F. W. Faltin (Eds.), *Encyclopedia of statistics in quality and reliability*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470061572.eqr089

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.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220-238.

Cassella, Robert A. (2010). Leader Development by Design. *ITEA Journal,* 31, 280-283.

Clark, R. E., & Estes, F. (1996). Cognitive task analysis for training. *International Journal of Educational Research*, *25*(5), 403-417.

Debeltz, R. (2017). *Engagement skills trainer: The commander's perspective*. US Army Command and General Staff College Fort Leavenworth United States.

Ericsson, K. A. (2009). Enhancing the Development of Professional Performance: Implications from the Study of Deliberate Practice. In *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments.* Cambridge University Press.

Goldberg, B., Owens, K., Hellman, K., Robson, R., Blake-Plock, S., Hoffman, M., Gupton, K. (2021). Forging Competency and Proficiency through the Synthetic Training Environment with an Experiential Learning for Readiness Strategy. In *Proceedings of the 2021 I/ITSEC*. Orlando, FL.

Goldberg, B., Hoffman, M. & Graesser, A. (2020). Adding a Human to the Adaptive Instructional System Loop: Integrating GIFT and Battle Space Visualization. In A. Graesser, X. Hu, Rus, V. and B. Goldberg (Eds.) Design Recommendations for Intelligent Tutoring Systems, Vol. 7: Data Visualization, U.S. Army Combat Capability Development Command.

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.

Klein, G. A., & Hoffman, R. R. (1992). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.), *Cognitive Science Foundations of Instruction* (pp. 203-226). Mahwah, NJ: Erlbaum.

Modeling and Simulation Glossary. (2011). Modeling and Simulation Coordination Office – United States Department of Defense. Retrieved June 10, 2022, from https://www.acqnotes.com/Attachments/DoD M&S Glossary 1 Oct 2011.pdf.

Ochoa, X., Lang, A. C., & Siemens, G. (2017). Multimodal learning analytics. *The handbook of learning analytics*, *1*, 129-141.

Perruchet, P. (1989). The effect of spaced practice on explicit and implicit memory. *British Journal of Psychology*, *80*(1), 113-130.

Pierce, L. V. (2002). Performance-Based Assessment: Promoting Achievement for Language Learners. Center for Applied Linguists (ERIC/CLL News Bulletin), 26(1), 1-3.

Sleeman IV, W. C., Kapoor, R., & Ghosh, P. (2021). Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *arXiv preprint arXiv:2109.09020*.

Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, *17*(2), 77-88.

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.

Synthetic Training Environment (STE). (n.d.). US Army Acquisition Support Center. Retrieved June 10, 2022, from https://asc.army.mil/web/portfolio-item/synthetic-training-environment-ste/

Tosuncuoglu, I. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies*, *6*(9), 163-167.

Vatral, C., Mohammed, N., Biswas, G., & Goldberg, B. S. (2021). GIFT External Assessment Engine for Analyzing Individual and Team Performance for Dismounted Battle Drills. In *Proceedings of the Ninth Annual Generalized Intelligent Framework for Tutoring Users Symposium (GIFTSym9)* (pp. 107-127). US Army DEVCOM – Soldier Center. (ISBN 13: 978-0-9977257-9-7).

Vatral, C., Biswas, G., & Goldberg, B. S. (2022a). Multimodal Learning analytics using hierarchical models for analyzing team performance. In *Proceedings of the 15th International Conference in Computer Supported Collaborative Learning* (In Press).

Vatral, C. Mohammed, N., Biswas, G., & Goldberg, B.S. (2022b). Moving Beyond Training Doctrine to Explainable Evaluations of Teamwork using Distributed Cognition. In *Proceedings of the Tenth Annual Generalized Intelligent Framework for Tutoring Users Symposium (GIFTSym10)* (pp. 127-137). US Army DEVCOM – Soldier Center. (ISBN 13: 978-0-9977258-2-7).

Zachary, W. W., Ryder, J. M., & Hicinbothom, J. H. (2000). Building cognitive task analyses and models of a decision-making team in a complex real-time environment. *Cognitive task analysis*, 365-384.

Zhang, N., Biswas, G., & Hutchins, N. (2021). Measuring and Analyzing Students' Strategic Learning Behaviors in Open-Ended Learning Environments. *International Journal of Artificial Intelligence in Education*, 1-40.