A Framework for Performance Assessment Across Multiple Training Scenarios Using Hierarchical Bayesian Competency Models

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 combines cognitive task analysis and expert input to design and develop a framework for assessing learner competencies and performance across multiple training scenarios. We adopt a hierarchical Bayesian approach to aggregate information from multiple modalities to derive competency metrics that relate to team coordination and individual psychomotor, cognitive, and affective measures of performance. The unified framework is represented as a task model that maps onto multiple task domains. The resulting hierarchical competency structure connects observed low-level performance measures for each task domain into higher-level competencies that are common across domains. By utilizing Bayesian inference to propagate evidence up the competency model, our framework is able to build a common model of high-level learner cognitive and psychomotor performance using evidence from multiple independent tasks. We demonstrate the effectiveness of the proposed framework using a case study of groups of soldiers performing two dismounted battle drills and show that the performance displayed by the soldiers provides consistent evidence for their higher-level competency states. With continued research and development, the proposed framework could allow for consistent longitudinal assessment of trainees based on observable evidence across a wide variety of domain skills and tasks.

ABOUT THE AUTHORS

Caleb Vatral is a PhD student at Vanderbilt University in the Department of Computer Science with a focus in intelligent systems. Working at the Institute for Software Integrated Systems, their research focuses on combining theoretical foundations in distributed cognition with multimodal data-driven approaches to support cognitive modeling and system design in simulation-based training and competency-based experiential learning. Prior to attending Vanderbilt, they 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.

A Framework for Performance Assessment Across Multiple Training Scenarios Using Hierarchical Bayesian Competency Models

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

Competency-based experiential learning has recently gained significant attention as an effective approach for training competencies and monitoring performance across a diverse set of training tasks. By focusing on the development and mastery of specific skills and abilities, this approach provides a comprehensive foundation for building and evaluating learners' abilities in various domains (Gervais, 2016). Understanding and assessing learner competencies is crucial for organizations and industries to ensure effective training and evaluate the readiness of individuals for specific roles or tasks. However, accurate assessment of learning competency is a difficult task, and current assessment approaches often face significant challenges in accurately capturing and evaluating learner behavior over the course of a training program. Traditional methods often rely on subjective observations, self-reporting, or summative assessments that may not provide a fine-grained analysis and a comprehensive understanding of a learner's true capabilities. Moreover, these approaches often lack consistency and fail to capture the complexity of performance across diverse tasks and domains. Instead of evaluating performance holistically across multiple tasks and domains, these approaches tend to evaluate performance on each task individually.

To address these challenges, there is a growing need for a comprehensive framework that enables consistent evidence-based assessment of high-level competencies across multiple tasks and domains. Such a framework will allow for a more holistic evaluation of learner competencies, taking into account various modalities such as team coordination, individual psychomotor skills, cognitive abilities, and affective measures of performance. By integrating multiple sources of evidence across a set of various tasks, we can develop a comprehensive understanding of learners' capabilities, enabling more informed decisions regarding their training and development.

Building and generalizing on our previous work, in this paper, we expand upon our novel approach that combines and extends cognitive task analysis, multimodal learning analytics, and theories of competency-based education to develop a framework for consistent data-driven performance assessment. Specifically in this work, we work toward solving the challenges of generalizing our developed hierarchical competency models across multiple training scenarios. Our approach adopts multimodal learning analytics, which allows us to aggregate information from multiple modalities and tasks to drive calculation of evidence-based competency metrics. Then, by mapping these metrics onto a unified task model that encompasses multiple tasks and domains, we establish a hierarchical competency structure that connects directly observable low-level performance measures to common higher-level competencies. We utilize Bayesian inference to enable the propagation of evidence up the hierarchical competency model. This process allows us to build a common model of high-level learner cognitive and psychomotor performance using evidence from multiple independent tasks. By leveraging this framework, we aim to provide a consistent longitudinal assessment of trainees based on observable evidence across a wide variety of domain skills and tasks.

To demonstrate the effectiveness of our proposed framework, we present a case study focused on the performance of groups of soldiers in two dismounted battle drills. The results of the study highlight the framework's ability to generate consistent evidence for the higher-level competency states of the soldiers across the independent drills. By addressing the challenges of current assessment approaches and presenting a comprehensive assessment framework, this research contributes to the advancement of competency assessment methodologies and opens avenues for further research and development in this field. We believe that with continued research and development, our framework holds the potential

to significantly impact and aid in the assessment of learner competencies, providing valuable insights into the capabilities of individuals and teams in various training scenarios.

BACKGROUND

Competency-Based Education

Competency-based experiential learning is a pedagogical approach that goes beyond traditional didactic education by emphasizing the development and mastery of specific competencies or skills (Gervais, 2016). This approach recognizes that true subject mastery involves not only theoretical knowledge but also the ability to apply that knowledge effectively in practical situations. By focusing on the development of competencies through experiences that mimic the true nature of the task environment, learners are better equipped to succeed in real-world scenarios and professional settings.

Competencies encompass a range of capabilities, including psychomotor skills, cognitive and metacognitive abilities, teamwork and communication, affective attributes, etc. When we model such competencies, we often breakdown and categorize them into either domain-specific or transferable (Nägele & Stalder, 2017). Domain-specific competencies, also known as low-level competencies in our previous work (Vatral et al., 2022a; Vatral et al., 2022b), refer to the specific knowledge and skills required to perform tasks within a particular domain. These domain-specific competencies are often highly specialized, but some examples could include the specific movement patterns of soldiers in battle drills, or the techniques, procedures, and algorithms used by medical staff in a hospital. Transferable competencies, also known as high-level or domain-general competences, on the other hand, are higher-level skills which are often applicable across various domains. Some examples include critical thinking, problem-solving, communication, teamwork, and adaptability. By emphasizing the development of both domain-specific and transferrable competencies and the interplay between them, competency-based experiential learning aims to prepare learners for the complexities and challenges of their chosen fields.

Cognitive Task Analysis

Cognitive Task Analysis (CTA) is a systematic and structured approach to understanding the cognitive processes and skills involved in performing tasks (Clark & Estes, 1996; Zachary, et al., 2000). It provides insights into the mental activities, decision-making processes, and problem-solving strategies that individuals employ while engaging in a task. By breaking down complex tasks into smaller components, CTA helps identify the knowledge, skills, and competencies required for successful task execution.

When performing CTA, we construct the models by combining detailed observation and review of the taskwork, interviews and close collaboration with domain experts, and careful review of relevant literature and doctrinal best practices. The models are structured hierarchically; each task is broken down into its constituent subtasks and those subtasks are in turn further broken down into more fine-grained subtasks. This iterative breakdown is repeated until the leaves of the hierarchy capture directly observable actions and behaviors. By analyzing the leaves of the CTA model and following the links up the hierarchy, we can gain insights about learners' higher-level cognitive behaviors by linking them to their observable actions.

Hierarchical Competency Modeling

In previous work (Vatral et al., 2022a; Vatral et al., 2022b), we developed a methodology for modeling learner competency at multiple levels of abstraction by creating a hierarchical competency model (HCM) based on a cognitive task analysis of the training domain. After performing CTA, we create the HCM as a parallel to the CTA model, with a similar hierarchical structure. At the highest levels of the model, the cognitive tasks contained in the CTA model are carried over and converted to parallel transferable competencies at the highest levels of the HCM. Since these concepts are already domain-general in the CTA model, they require little transformation to convert them to parallel transferable competencies. This parallelization process then continues as we move down the CTA hierarchy, with each layer moving from more domain-general transferable competencies to more domain-specific competencies. At the lowest levels of the CTA model, we have highly-domain specific actions that are converted to highly domain-specific metrics that measure performance on those actions in the parallel HCM. An example HCM, which we call the H-ABC model (Vatral et al., 2022a), is illustrated in figure 1 and demonstrates this hierarchical structure with high-level transferable

competencies, i.e., teamwork, cognitive, and behavioral, at the top of the hierarchical, and each subsequent layer presenting more and more objective domain-specific concepts, until we reach the very bottom of the hierarchy, which contains domain-specific metrics.

After defining the HCM conceptually in parallel to the CTA model, we define it mathematically as well, which will allow us to make inferences about unobservable higher-level transferable competencies using the observable domain-specific competencies to which they are connected. First, these low-level domain-specific competencies are defined using domain-specific performance metrics, as previously mentioned. The exact method of computing these metrics can be determined by the authors of the HCM depending on what learner data is available, but in previous work and this work, we use computer vision analysis on recorded video (Vatral et al., 2022b).

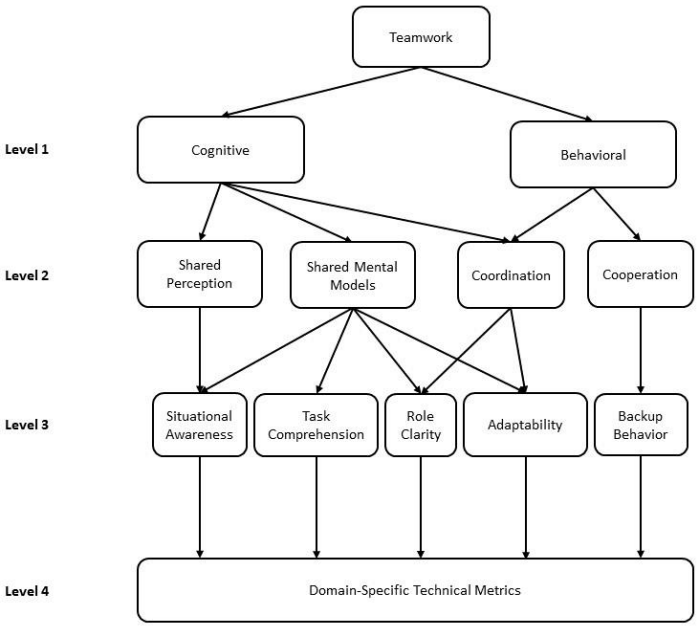

Figure 1. The H-ABC hierarchical competency model used for this study

After defining these metrics, we then define how they relate to the higher-level transferable competencies at each layer above them. Conceptually, this relationship is modeled by the hierarchical links in the CTA/HCM. Mathematically, we define these relationships using a Bayesian network, which is a state-based graphical probability model that allows for inference about unobservable variables using connected evidence variables (Ben-Gal, 2008). We represent each node in the HCM as a variable in the Bayes net which can take on one of three states – below-expectation, at-expectation, or above-expectation – based on the three-state learner models employed both in Generalized Intelligent Framework for Tutoring (GIFT) (Goldberg, et al., 2021), where our system is implemented, as well as across the literature on training (Cassella, 2010; Klein & Hoffman, 1992; Sottilare, et al., 2017). For each node, we define a prior-probability distribution over these three states, which represents the initial probability that the trainees are in each of these states. In theory, this prior-probability could encode the prior experience level of a trainee or team; for example, a team could have high probability of being in the above-expectation state on the communication competency if they have trained together for a long time, while a new team that was just formed might have higher probability of being in the below-expectation state. In practice, we often initialize this prior probability to 100% below-expectation, as we often do not have a good initial assessment of the team's competency or information about how long they have trained together.

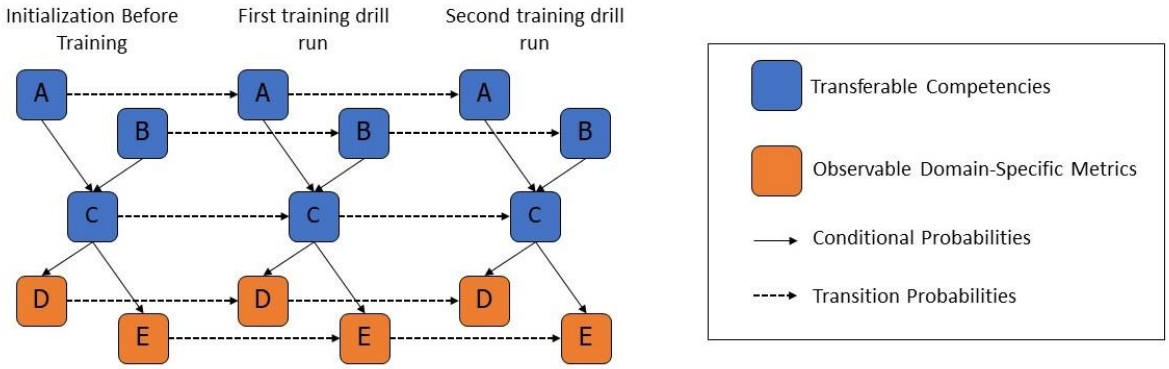

Figure 2. Illustration of the dynamic Bayesian network update procedure.

Finally, we define the links between nodes in the HCM as a Bayes net with an associated conditional probability distribution. This distribution encodes how two competencies are related to one another. For example, a higher-level concept such as *coordination* is more likely to be in the above-expectations state if its child competencies, *role clarity* and *adaptability* are in the at- or above-expectation states than if they are in the below-expectation state. The overall idea is that performance on a higher-level transferable competency is conditionally dependent on exhibiting performance on low-level domain-specific competencies. This allows us to infer the states of unobservable competencies given the evidence of low-level measurement competencies and metrics.

This process is illustrated in figure 2, which shows the relationships between observable domain-specific metrics, unobservable transferable competencies, and updates over time. First, the HCM is initialized using the prior-probability distributions. Next, we receive evidence in the form of domain-specific metrics from a training drill. The evidence at these domain-specific nodes at time *t* are used to infer the states of the transferable competencies at time *t* using Bayesian inference on the conditional probabilities and prior probabilities. Then, the newly inferences states become are used along with the transition probabilities to calculate the prior probabilities for inference at time *t+1* and the process repeats itself. For more details about the theoretical and mathematical formulation of our HCM, see Vatral, et al. (2022b) and Vatral, et al. (2023).

METHODS

Case Study Design

In this work, we build upon the case study examined in Vatral et al. (2022). In our study, a fire team of soldiers trained on two dismounted battle drills, *Enter and Clear a Room* and *Break Contact*, over the course of one full day. In the morning, the team trained on the *Break Contact* drill. In the afternoon, the team switched to the *Enter and Clear a Room* drill. All the training was completed using the Squad Advanced Marksmanship Trainer (SAM-T), which is a mixed-reality training platform. The SAM-T projects a VBS3 simulation onto a set of screens, and the soldiers then move around the physical space around the screens. Figure 3 shows the setup of each drill. Soldiers can interact with the simulation using modified weapons which are interfaced with the VBS simulation. In addition, instructors can modify the simulation in real-time based on any actions taken by the soldiers (verbal commands, movements, gestures, etc.). The SAM-T provides data of the simulations including video, audio, soldier biometrics, and simulation logs which were synchronized and recorded live using the GIFT data collector. For this study, we analyze the collected video data to derive performance evaluations.

Figure 3. The two dismounted battle drills run on the SAM-T used for this study. Top: Break Contact; Bottom: Enter and Clear a Room

Metric Design and Domain Adaptation

One of the primary goals of this work is to show how the methods presented in our previous work (Vatral et al., 2022) can be adapted and transferred to various different domains while supporting a common evaluation of higher-level transferable competencies. In this way, the case study presented here provides a good test case, as each drill trained and evaluated similar transferable psychomotor and teamwork skills, while having differing implementations at the domain-specific level. To capture this in the HCM, we build upon our extensible H-ABC model of teamwork as the basis for the HCM. The H-ABC model, shown in Figure 1, provides the first 3 layers of the competency model, which

evaluates the common transferable competencies for both drills. To complete the model, as described in the previous section, each drill then needs a set of performance metrics which are measurable from the collected trainee data. For each of the drills, we designed 5 performance metrics which capture the psychomotor performance of the soldiers as it relates to their overall teamwork. All of the metrics were calculated using computer vision techniques to track the soldiers' movements and psychomotor behaviors. For the *Enter and Clear a Room* drill, we reused the same 5 metrics which appeared in Vatral et al. (2022). For the *Break Contact* drill, we designed 5 new metrics for this analysis which can be computed from the same motion tracking algorithms used in the previous study. The metrics for both drills are outlined in Table 1 and Table 2 and their connections to the H-ABC model are shown in figure 4.

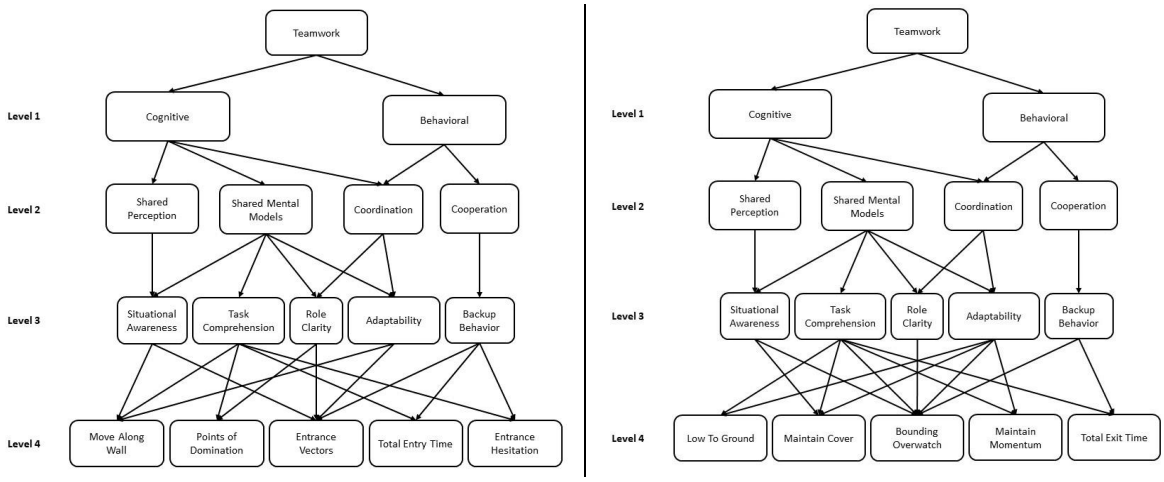

Figure 4. The H-ABC model with connected domain-specific metrics for the *Enter and Clear a Room* drill (Left) and the *Break Contact* drill (Right).

In order to make inference about the transferable competencies using both sets of domain-specific metrics, we primarily follow the same Dynamic Bayes Network inference procedure as presented in Vatral et al. (2022) and explained in earlier sections but make some small modifications to allow for both domains to influence the high levels independently. First, the Bayes net is initialized with the prior probabilities of the team. As previously mentioned, for this study we initialize all competencies to *Below-Expectation* with 100% probability, as we do not have baseline data for how the teams start. Next, we connect the Bayes net to the domain-specific metrics for the first drill that a team performs for the day. Then, as they perform each repetition of the drill, we propagate the evidence variables up the Bayes net to update our beliefs about the learner state of each competency according to the prior probability from the last repetition, the conditional probabilities defined during the creation of the HCM, and the transition probabilities between states. This follows exactly the paradigm presented in the previous work. However, once the team switches to the next drill, we freeze the probabilities of the higher-level transferable competencies and switch out the lowest layer of the HCM to the domain-specific competencies for the new drill. Once switched, the dynamic Bayes net can continue to be updated with evidence obtained from subsequent drills.

Table 1. The five domain-specific performance metrics which were calculated for the *Enter and Clear a Room* drill.

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

Table 2. The five domain-specific performance metrics which were calculated for the *Break Contact* drill.

Metric Name	Description	Calculation
Bounding Overwatch	Is there at least one soldier maintaining cover fire while others are moving?	Percentage of time that at least one soldier is stationary while other soldiers are moving
Maintain Cover	Are the soldiers behind cover when they are not moving?	Percentage of time that soldiers spend behind a cover object when not moving
Maintain Momentum	Are the soldiers maintaining a consistent retreat from the enemy after making contact?	Percentage of segments when the entire team is stopped that last longer than an optimal threshold
Total Exit Time	How quickly does the team reach a safe distance after making contact?	Normalized difference between the team's exit time compared to the optimal time threshold
Low To Ground	Does the team maintain a posture close to the ground when not moving?	Percentage of time that the team is kneeling or prone when stopped

In more detail, at each timestep which corresponds to each drill execution, we collect a set of evidence variables, *E*, which in our case represent the observed domain-specific performance metric scores. Then, we inference about the posterior probability, $P(H|E)$, of each parent node of the performance metrics using the defined prior probabilities, $P(E)$, the defined conditional probabilities, $P(E|H)$, and the marginal likelihood of the observed evidence, $P(E)$. This is the standard Bayesian network update procedure based on Bayes' theorem. However, in our case, we additionally have information about the history of the trainee performance, which we encode in our model by an additional conditional dependency, called the transition probability (see figure 2 for an illustration). This formulates the Bayesian network into a dynamic Bayesian network. For a simple example, consider the *Role Clarity (R)* competency under the *Break Contact* drill, which has one connected child metric, *Bounding Overwatch (B).* In this case, we can calculate the posterior probability state of *R* at time *t*, $P(R_t \mid B_t. R_{t-1})$, based on the prior probability, $P(R)$, the conditional probability of the evidence, $P(B_t \mid R_t)$, the conditional probability of the prior time step, $P(R_t \mid R_{t-1})$, and the marginal probability of the observed evidence, $P(E)$, as shown in equation 1. This process and calculation are similar for other competencies as well, with additional terms for each connected child.

$$P(R_t \mid B_t, R_{t-1}) = \frac{P(B_t \mid R_t) \cdot P(R_t \mid R_{t-1}) \cdot P(R)}{P(B_t)} \quad (1)$$

For our case study, we initialize the prior probability at time *t=0* to 100% below-expectation and use a set of hand-designed conditional and transition probability models shown in table 3. The overall idea of the hand-designed conditional model is that parent and child concepts in the H-ABC hierarchy are likely to be in the same state, as understanding and mastery of one concept relies on understanding and mastery of the other. In cases where a child has two or more parents, the full conditional probability distribution is constructed by multiplying the conditional probability distributions of each of its parents. This overall idea of the hand-designed transition model is that a trainee will very likely not transition between competency states after any single training instance, but by performing multiple training instances back-to-back (as in our case-study), the probability of transitioning adds up.

Table 3. The hand-designed probability models used in the H-ABC Bayesian network for our case-study.

(a) Conditional Model

	Below	**At**	**Above**
Below	0.75	0.2	0.05
At	0.2	0.6	0.2
Above	0.05	0.2	0.75

(b) Transition Model

	Below	**At**	**Above**
Below	0.95	0.05	0
At	0	0.95	0.05
Above	0	0	1

RESULTS

Using the calculated domain-specific metrics as the evidence variables and the Bayesian propagation procedure previously outlined, we were able to infer the learner states of the higher-level transferable competencies across both drills for the entire day of training. Figure 5 shows the performance progression of the fire team. 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 these results, two key ideas emerge.

First, the patterns and themes identified in the previous analysis in Vatral et al. (2022) largely still hold true. We still see that the team improves in most competencies as the day progresses; we still see that the team mastered lower-level domain-specific concepts first, which then lead to their mastery over higher level transferable concepts; and we still see that state transitions for transferable competencies are smoother than transitions for domain-specific competencies. The fact that these same patterns emerged is not entirely surprising, since some of the same data was used for both analyses; however, this does serve to demonstrate that the advantages and key features of the methods we have developed – i.e., monitoring of performance progression, mathematical structures mimicking expected human performance, and plateaus in high-level competencies to demonstrate learning saturation, etc. – still maintain their validity when examining multiple drills.

Second, and perhaps more importantly, the patterns we see in the data strongly support the notion that in order to fully understand learners' mastery of transferable competencies, we have to analyze the learners in multiple training contexts. This conclusion is most evident when we compare the performance progression analysis generated using only a single drill to the analysis generated using both drills. To demonstrate this, consider the *Situational Awareness* competency. Figure 6 (top) shows the probability that the team is in the *at-expectation* state for the *Situational Awareness* competency when we run our Bayesian analysis using only the data from the *Enter and Clear a Room* drill. From this data, we can see a noisy but clear trend upward (R2 = 0.61), indicating that the team is improving on this competency as the day progresses. However, we can contrast this to the Bayesian analysis using data from both drills, as shown in Figure 6 (bottom), which tells us a very different story. When using both drills, we can see a quick and clear trend upward in their competency (R2 = 0.91) when performing the initial *Break Contact* drill. However, once the team switches to the *Enter and Clear a Room* drill, their performance quickly drops off and then follows with only a small trend upward (R2 = 0.14) for the rest of the day.

Taken together, this suggests that when each drill is analyzed individually, we see strong learner progress, but when both drills are analyzed together, we see that the *Situational Awareness* concepts that the team learned demonstrated mastery of during the *Break Contact* drill did not transfer well to the *Enter and Clear a Room* drill. It was only once the model saw the performance of the team on the *Break Contact* drill that the true performance trend for the *Enter and Clear a Room* drill could be captured. Without evidence from both drills, we would not be aware of this skill transfer issue and might be lured into a false sense of security about the true competence of the team's *Situational Awareness.* Performance evidence from both drills is necessary to fully demonstrate and understand the team's competency.

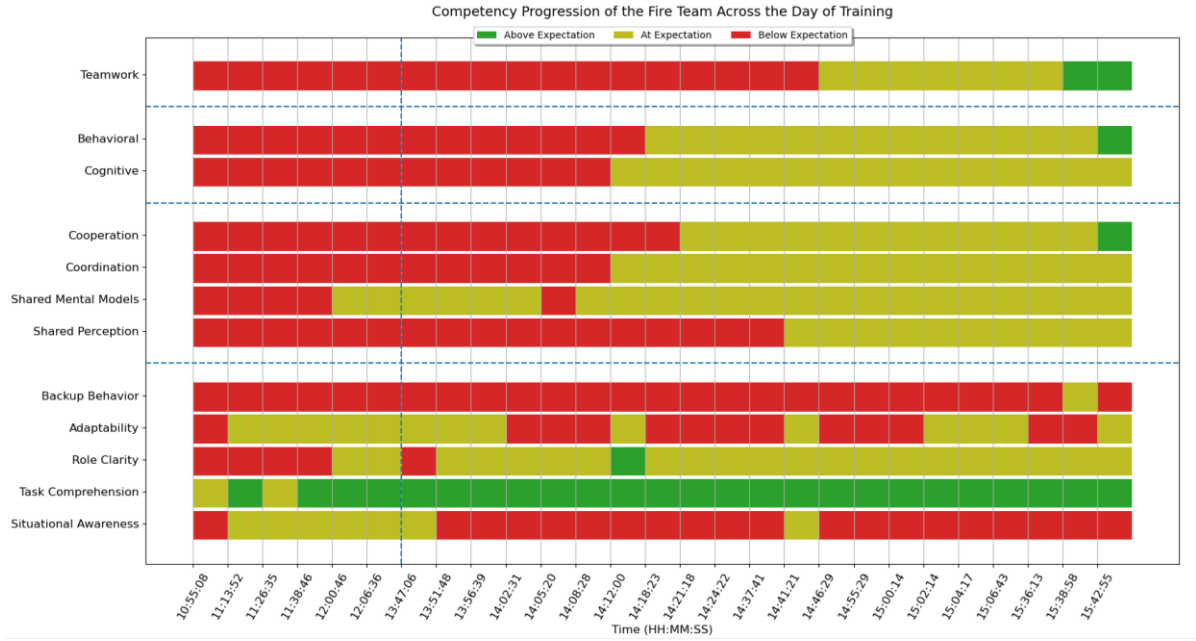

Figure 5. Performance progression across each level of abstraction in the H-ABC model over the course of the entire training day.

DISCUSSION

In this section, we discuss the broader implications of our automated assessment framework for competency-based training programs. Based on the results of this study, there are several key takeaways: (1) Lessons learned during domain adaptation of our methods; (2) highlighting the transfer of skills between multiple training domains; (3) the importance of integrating multiple training scenarios into data-driven assessments in competency-based education training programs.

First, there were several lessons learned during the domain adaptation process which allowed us to analyze both of these drills using a consistent framework. The first challenge of this process was generating new metrics for the *Break Contact* drill that were capable of assessing performance of the same high-level transferable competencies as the *Enter and Clear a Room* drill. In our case, this process was aided by many psychomotor similarities between the battle drills, but in general, the further apart the drills are from one another, the more difficult this process would become. This highlights the importance of starting with a flexible hierarchical competency model which is capable of capturing a wide variety of individual and team behaviors, such as the H-ABC model used here. Then, by designing new metrics for the new drill and connecting them to a comprehensive model, it is easy to identify areas where the current set of

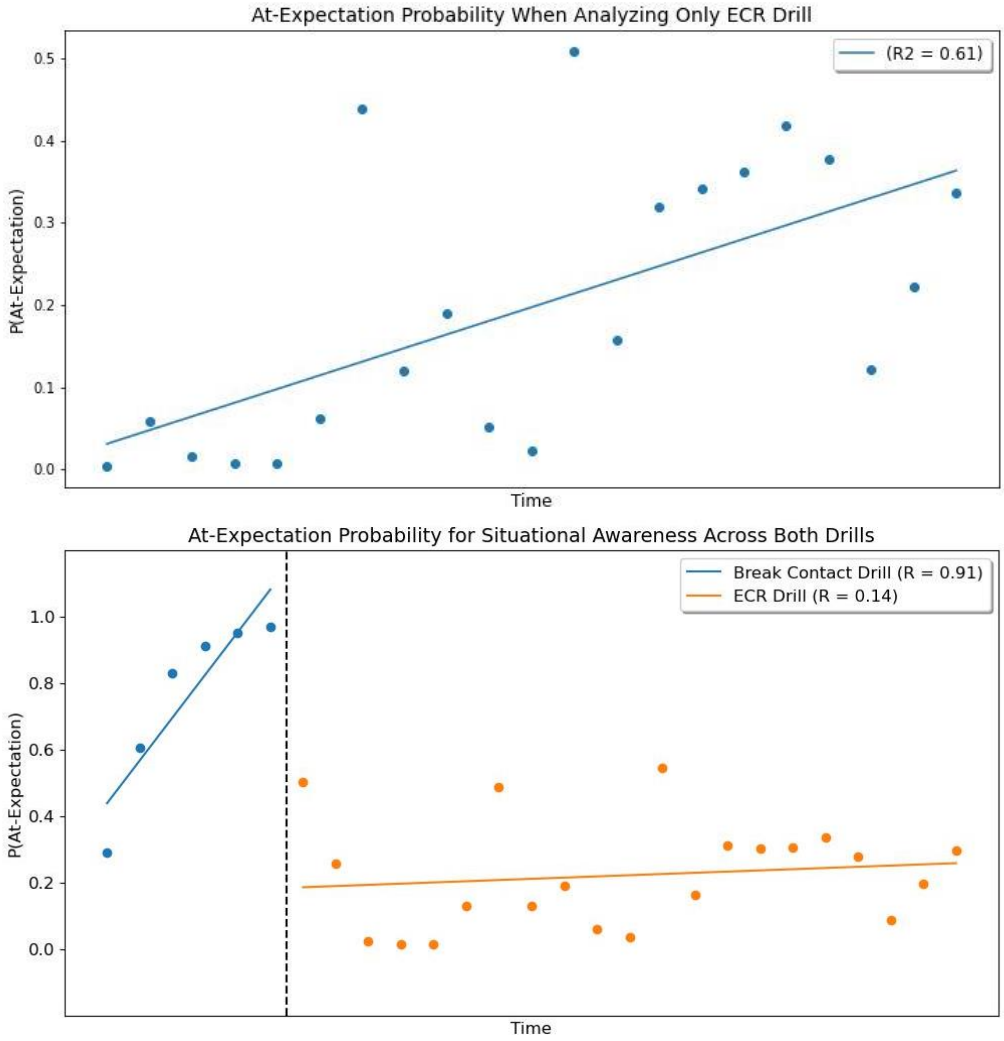

Figure 6. Probability of the team being in the *at-expectation* state for the *Situational Awareness* competency when analyzed using only the ECR data (top) and when analyzed using data from both drills (bottom).

metrics are lacking assessment. Second, after these metrics are designed, there are additional challenges with adapting the algorithms used to calculate these metrics. In our case, we had to overhaul and fine-tune the computer vision-based tracking algorithms to handle edge cases which appeared in the new drill but not the old drill. Thus, when designing assessment algorithms, it is important to be mindful of the extensibility of the methods and ensure that assessments make use of as much domain-general tooling as possible to limit the amount that they must be redesigned or retrained when adapting to new domains.

Second, the results of the case study, especially when compared to previous studies, highlight the capabilities of our hierarchical Bayesian modeling to generate insights about how different skills transfer between different domains. As we saw in the previous section, when the team switched between drills, several competencies saw immediate and sustained drops in performance. This was quintessentially exemplified through the *Situational Awareness* competency that we analyzed in depth but was also a common pattern in several other competencies. When implementing a training program, it is very important for instructors to recognize these drops in performance between drills, as it indicates a larger problem of lack of transfer between the two drills. By recognizing these situations and responding appropriately, instructors can highlight similarities where transfer should occur to trainees in order to maximize the cross-domain learning gains. Data-driven automated assessment frameworks, such as ours, provide a valuable tool to help recognize and point out these transfer gaps to instructors so that they can address them with the learners.

Finally, the results of the case study significantly underscore the importance of training across a diverse spectrum of exercises and conditions to adequately evaluate the competency of an individual or team. While this is not a brand-new idea in the competency-based education and training space, our case-study serves as an important demonstration of this idea empirically. Automated assessment models such as ours are limited by the data evidence which they are given, but if that evidence has significant unforeseen gaps, such as the lack of knowledge transfer that we saw in the previous section, then the models may be overly optimistic about their estimates of the true competency of the learners. It is only by integrating evidence from multiple training drills and training conditions that support assessment of the same transferable competencies that we understand the full breadth of a learner's true competence.

CONCLUSIONS AND FUTURE WORK

In this paper, we extended our hierarchical Bayesian automated performance assessment system for competency-based experiential learning to allow for multiple diverse training drills to simultaneously evidence the same higher-level transferable competencies. By examining a case study of a fire team of soldiers training on two dismounted battle drills, we showed the cross-domain extensibility of our methods and highlighted several key takeaways from implementation and practice including lessons learned for domain adaptation, the capabilities of the system for evaluating knowledge transfer, and empirical evidence for the importance of evaluating competency using multiple diverse training scenarios. Though the results presented here are promising, this study had several limitations which should be addressed in future research. First, this paper focused on a case-study evaluating one team across only two drills. Future work should repeat this modeling and evaluation with more teams and across additional training scenarios in order to further validate the results. Furthermore, the analysis presented in this paper was performed offline post-hoc, not in real-time during active training. This approach was necessary during the development phase as it facilitated rapid iteration on prototypes without the need to schedule additional studies. However, future studies should implement this assessment framework live during the training and data collection, which would allow the assessments to be presented back to instructors and trainees for feedback to ensure its suitability for the system's intended audience. With continued development, we hope that our hierarchical Bayesian competency assessment framework can be a valuable addition to a larger training management tool and help to improve the learning outcomes of its users.

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. 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

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

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.

Gervais, J. (2016). The operational definition of competency-based education. *The Journal of Competency-Based Education*, 1(2), 98– 106. doi: 10.1002/cbe2.1011

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.

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

Nägele, C., Stalder, B.E. (2017). Competence and the Need for Transferable Skills. In: Mulder, M. (eds) Competence-based Vocational and Professional Education. Technical and Vocational Education and Training: Issues, Concerns and Prospects, vol 23. Springer, Cham. https://doi.org/10.1007/978-3-319-41713-4_34

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.

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* (pp. 403-406). International Society of the Learning Sciences.

Vatral, C., Mohammed, N., Biswas, G., & Goldberg, B. S. (2022b). Multimodal Learning analytics using hierarchical models for analyzing team performance. In *Proceedings of the 2022 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. National Training and Simulation Association.

Vatral, C., Mohammed, N., Biswas, G., & Goldberg, B. S. (2023). A Theoretical Framework for Multimodal Learner Modeling and Performance Analysis in Experiential Learning Environments. In *Workshops of the 2023 International Conferenece on Aritificial Intelligence in Education (AIED)* (In Press). International Society of Artificial Intelligence in Education.

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.