Advancing Squad Performance Analytics and Team Training with Multimodal Data in STEEL-R

Randall Spain, Benjamin Goldberg, Lisa Townsend
U.S. Army DEVCOM Soldier Center
Orlando, Florida
randall.d.spain.civ@army.mil
benjamin.s.goldberg.civ@army.mil
lisa.n.townsend2.civ@army.mil

Grace Teo
Quantum Improvements Consulting
Orlando, Florida
gteo@quantumimprovements.net

Nicholas Roberts
Dignitas Technologies
Orlando, Florida
nroberts@dignitastechnologies.com

Clifford Hancock, Meghan O'Donovan
U.S. Army DEVCOM Soldier Center
Natick, Massachusetts
clifford.l.hancock4.civ@army.mil
meghan.p.odonovan.civ@army.mil

ABSTRACT

Advancements in virtual and live training technologies are providing new opportunities to integrate multimodal data—video, audio, sensor, geographical positioning, and behavioral data—into training evaluations to provide data-enhanced assessments. Using multi-modal data to assess team performance and inform learning analytics is a critical dependency for providing squads and Soldiers with timely feedback, which is a key focus of the U.S. Army's training modernization program. Achieving this end-state requires a robust data infrastructure to reliably capture, synchronize, and interpret multimodal data sources, an evidence-centered framework for converting raw data into actionable insights that can support performance assessment, and functions to support data visualizations and performance modeling. Over the past year, researchers at the U.S. Army DEVCOM Soldier Center have focused on enhancing the Synthetic Training Environment Experiential Learning for Readiness (STEEL-R) architecture to address these challenges. By collecting diverse data sources—including motion capture, GPS, real-time communication logs, physiological responses, and fire effectiveness data—and integrating these sources into a squad competency framework, STEEL-R aims to provide faster and more accurately aligned insights into the factors and processes impacting team dynamics, individual behaviors, and small unit performance. This paper discusses how we have systematically extended the STEEL-R architecture to interoperate with a state-of-the-art mesh network to capture live instrumented training data and align it to a squad competency framework in near real time. We outline the architectural requirements for accommodating multimodal data streams, discuss an assessment model we developed that transforms raw data into insights for competency modeling, and describe enhancements to STEEL-R's data visualization capabilities to improve after-action review experiences. Finally, we highlight case study outcomes from applying STEEL-R to assess squad performance during a field training exercise.

ABOUT THE AUTHORS

Randall Spain, PhD., is a Research Scientist at the U.S. Army Combat Capability Development Command – Soldier Center. He holds a Ph.D. in Human Factors Psychology from Old Dominion University and an M.S. in Experimental Psychology. His research focuses on designing and evaluating adaptive and intelligent training systems.

Benjamin Goldberg, **PhD.,** is a Senior Research Scientist at the U.S. Army Combat Capability Development Command – Soldier Center. His research focuses on adaptive experiential learning with an emphasis on simulation-based environments and leveraging Artificial Intelligence to create personalized experiences. Dr. Goldberg holds a Ph.D. in Modeling & Simulation from the University of Central Florida and is well published across several high-impact journals and proceedings, including IEEE Transactions of Learning Technologies, the Journal of Artificial Intelligence in Education, and Computers in Human Behavior.

Lisa N. Townsend, M.S., is a Senior Research Psychologist at the U.S. Army Combat Capabilities Development Command Soldier Center, Simulation & Training Technology Center. She has an M.S. in Industrial/Organizational Psychology and a B.A. in Psychology, from the University of Central Florida (UCF). She has worked on many diverse

teams including those within Research and Development, Technology Transfer, Instructional Systems Design, and Human Systems Integration. Ms. Townsend's areas of expertise involve team training, Front End Analysis (FEAs), Training Systems Analyses (TSAs), Instructional Systems Design (ISD), Training Effectiveness Evaluations (TEEs), and the development of training and organization related metrics. Her efforts in these areas have spanned across Services and platforms.

Grace Teo, Ph.D., is Lead Learning Scientist at Quantum Improvements Consulting. Grace's research involves understanding and improving human performance under various conditions and in different contexts, such as working with different technologies, and in teams. Other research interests include assessments, decision-making processes and measures, vigilance performance, human-robot teaming, automation, and individual differences. Grace earned her Ph.D. and M.A. in Applied Experimental and Human Factors Psychology from the University of Central Florida.

Nicholas Roberts, a Senior Software Engineer at Dignitas Technologies and the engineering lead for the GIFT project, has been involved in the engineering of GIFT and supported collaboration and research with the intelligent tutoring system (ITS) community for nearly 10 years. Nicholas contributes to the GIFT community by maintaining the GIFT portal (www.GIFTTutoring.org) and GIFT Cloud (cloud.gifttutoring.org), supporting conferences such as the GIFT Symposium, and technical exchanges with Soldier Center and their contractors.

Clifford Hancock, M.S., is a biomechanical engineering researcher with the U.S. Army Combat Capabilities Development Command – Soldier Center (Natick, MA, USA). He received his M.S. in Biomedical Engineering from the University of Texas at Arlington in 2013 and his B.S. in Engineering Science and Mechanics from the Virginia Polytechnic Institute and State University (Virginia Tech) in 2011.

Meghan O'Donovan, M.S., is a principal investigator and project officer with the U.S. Army Combat Capabilities Development Command – Soldier Center (Natick, MA, USA) who has led several efforts in the areas of human performance, military equipment evaluation, and performance augmentation. She currently leads the U.S. Army Small Unit Performance Analytics program (SUPRA). She holds an M.S. in Biomedical Engineering with a concentration in Biomechanics from the University of Rochester.

Advancing Squad Performance Analytics and Team Training with Multimodal Data in STEEL-R

Randall Spain, Benjamin Goldberg, Lisa Townsend
U.S. Army DEVCOM Soldier Center
Orlando, Florida
randall.d.spain.civ@army.mil
benjamin.s.goldberg.civ@army.mil
lisa.n.townsend2.civ@army.mil

Grace Teo
Quantum Improvements Consulting
Orlando, Florida
gteo@quantumimprovements.net

Nicholas Roberts
Dignitas Technologies
Orlando, Florida
nroberts@dignitastechnologies.com

Clifford Hancock, Meghan O'Donovan
U.S. Army DEVCOM Soldier Center
Natick, Massachusetts
clifford.l.hancock4.civ@army.mil
meghan.p.odonovan.civ@army.mil

INTRODUCTION

The U.S. Army, like other professional organizations, is investigating how to leverage multiple data sources to assess and optimize Soldier performance. By instrumenting Soldiers with biometric sensors, microphones, cameras, and additional data collection devices the Army aims to assess, predict, and enhance Soldier lethality and small unit effectiveness through data-driven decision-aids and targeted interventions. One of the challenges of collecting these forms of multimodal data is determining how best to store, manage, and use these data in a manner to support assessing and modeling Soldiers performance. Another challenge is identifying best practices for combining performance measures from synthetic (i.e., virtual) and live training events to support competency modeling.

Competency modeling is a critical function of intelligent and adaptive training systems, providing information about learner attributes, knowledge, and proficiency states that can be used to adapt training experiences. Using multimodal data sources to support competency modeling and performance visualization aligns with U.S. Army's modernization efforts to use intelligent tutoring and after-action review (AAR) technologies to accelerate learning and reduce training costs (U.S. Army, 2019). Effective data visualizations used in AARs can prompt Soldiers to engage in reflection, peer comparison, and discussion to improve collective performance (Johnston et al., 2020; Salas et al., 2018). Leveraging multimodal data to support these forms of enhanced AARs experiences requires a capable data infrastructure that can collect, process, assess, and store performance data in near real time.

Over the past three years, the U.S. Army Combat Capabilities Development Command (DEVCOM) Soldier Center has been leading research to integrate multimodal data from live and virtual training events into a scalable competency management framework using the STE-Experiential Learning for Readiness (STEEL-R) architecture. STEEL-R is a Science and Technology investment that supports the U.S. Army's training modernization effort (Goldberg et al., 2021; Hernandez et al., 2022). It uses a suite of open-source software applications to create a persistent data ecosystem capable of capturing and logging granular metrics of performance across time. These metrics are securely stored in a learning record store and competency model that computes estimates of individual and team competency states (Robson et. al. 2022). The data strategy implemented in STEEL-R is guided by the theory of experiential learning (Kolb & Kolb, 2017) which posits that skill mastery and proficiency require repeated, deliberate practice under varied conditions, and emphasizes the importance of reflection and conceptualization within the learning cycle. STEEL-R's underlying data strategy has been validated with simulated human performance and training data (Robson et al., 2022), however, applying this approach in a real-world setting is critical to maturing its competency modeling methods.

This paper presents a case study detailing the systematic integration of data from a live field training event into STEEL-R's data pipeline and competency management framework, thereby advancing its capabilities. Specifically, we describe how we extended STEEL-R's architecture to interoperate with a state-of-the-art mesh network to ingest multimodal data from instrumented infantry squad members performing Battle Drill 2A (BD2A) "React to Contact". We discuss the alignment of this data to a squad competency framework and how we used STEEL-R's components to assess squad performance in near real-time. Furthermore, we present results examining the impact of STEEL-R's

enhanced AAR capabilities (facilitated by data visualization views) on perceived AAR effectiveness ratings gathered from squad members. This paper makes three primary contributions:

1. It extends the STEEL-R architecture to support integration of multimodal sensors and observational data from live infantry training exercises.
2. It discusses a competency-based assessment framework that maps behavioral measures to teamwork constructs in a hierarchical competency model, supporting automated squad performance evaluation.
3. It presents results showcasing how real-time performance visualizations can enhance reflective learning and AAR effectiveness.

In the sections that follow, we describe the theoretical foundations of this work, the design of the STEEL-R architecture, the deployment of the system in a live training context, and the resulting implications for Army training effectiveness and readiness.

Background and Theoretical Foundation

U.S. Army units invest significant time and resources to promote and maintain readiness. Through repeated training experiences Soldiers rehearse mission essential tasks and their corresponding cognitive and behavioral skills across repeated sets and reps to build mastery. When coupled with guided feedback and reflective AAR experiences, training experiences can significantly enhance individual and team performance (Keiser & Arthur, 2022; Salas et al., 2008). Traditional AARs often rely on subjective observer/controller (OC/T) ratings and static evaluation checklists, which can limit the granularity, objectivity, and consistency—especially in fast-paced small-unit operations (Johnson & Gonzalez, 2008). To provide objective and unobtrusive approaches for small unit assessment, Army modernization initiatives are incorporating data-centric approaches (Hancock et al., 2025). Multimodal data—including movement, physiological, audio, visual, and engagement data—offer new opportunities to understand, model, and accelerate learning (Sharma & Giannakos 2020; Vatral et al., 2022). Multimodal data can provide objective indicators of both individual and team performance, particularly when aligned to structured training objectives and competency frameworks.

STEEL-R is grounded in the principles of Experiential Learning Theory (ELT), which stresses the importance of learning by doing through Concrete Experiences, Reflective Observation, Abstract Conceptualization, and Active Experimentation (Goldberg et al., 2021; Kolb, 1984). This theory attests learning is most effective when individuals actively engage in a learning experience, reflect on that experience, draw insights, and then test those insights in new situations. The U.S. Army's training model closely aligns with this learning cycle. Units and Soldiers conduct training at dedicated cycles to learn, practice, and refine mission essential tasks. When a training exercise concludes, units reflect on their performance as a part of their AAR. Facilitators or Observer/Controllers and Trainers (OC/Ts) often guide these sessions by highlighting successes, identifying areas for improvement, and helping Soldiers extract lessons learned (Meliza & Goldberg, 2017). Soldiers and units can apply the insights gained from these reflective discussions in subsequent iterations of training, completing the learning cycle with active experimentation.

The AAR methodology fundamentally relies on observation to drive reflection and learning. However, even highly experienced Observers/Trainers (OC/Ts) may miss critical actions or factors impacting unit performance due to limitations in their vantage point. Over time, AAR methodologies have evolved to incorporate videos, simulation replays, and structured observations to address these limitations and improve effectiveness (Morrison & Meliza, 1999). Building on this evolution, technology-enhanced AARs that integrate multimodal performance data offer the potential to improve the objectivity of assessments and the quality of feedback. By presenting data-driven visualizations of squad behavior, these systems can deepen reflection, support peer-to-peer learning, and reinforce effective behaviors (Johnston et al., 2020).

Multimodal Learning Analytics

Multimodal learning analytics is an emerging area of research focused on capturing and interpreting complex data streams to understand learning processes and outcomes (Giannakos & Cukurova, 2023). Researchers have used multimodal data to assess different learning phenomena including collaborative learning (Acosta et al., 2024), conceptual and procedural knowledge (Ainam et al., 2025; Emerson et al., 2023), team performance (Smith et al., 2023) and self-regulated learning behaviors (Azevedo et al., 2024). Recent applications have extended multimodal

learning analytics to military training, where data from GPS, video, and communication systems have been used to assess small unit performance and identify exemplary unit coordination behaviors (Hancock et al., 2025; Vatral et al., 2022).

However, effectively interpreting these complex data streams requires a framework for defining and measuring the specific skills and behaviors that contribute to successful performance. To support this, researchers often turn to competency modeling. A competency model defines the knowledge, skills, and attributes required for successful task execution and links these actions and constructs to observable behaviors and performance metrics (Goldberg et al., 2021; Owens & Goldberg, 2022). Competency models can be integrated into training systems to provide formative and summative assessments, facilitating personalized feedback, skill tracking, and tailored training interventions.

The STEEL-R project directly operationalizes these principles within the U.S. Army's Synthetic Training Environment. Developed to support competency-based training and assessment, STEEL-R applies learning engineering principles to combine instructional design, sensor integration, and data science into a scalable, standards-based training architecture (Blake-Plock et al., 2023). By connecting training events to defined competency frameworks and leveraging real-time data capture, STEEL-R aims to enable persistent performance assessment across the Soldier's learning continuum, providing units with timely, data-guided, and contextually driven insights into the factors and processes that mediate unit performance (Goldberg et al., 2021).

STEEL-R ARCHITECTURE

STEEL-R is a modular, open-source data architecture designed to support real-time, competency-based assessment of Soldier and squad performance across live, virtual, and constructive training environments (Hernandez et al., 2022). Built on a Modular Open Systems Architecture (MOSA) design, STEEL-R integrates a suite of interoperable technologies that enable persistent data capture, adaptive instruction, and longitudinal competency modeling. Its design aligns with the U.S. Army Learning Model 2030 - 2040 (TRADOC, 2024) and broader modernization objectives of the Total Learning Architecture (TLA) that emphasize intelligent tutoring systems and data-informed decision-making (Goldberg et al., 2021; Owens et al., 2020).

A central component of STEEL-R is the Generalized Intelligent Framework for Tutoring (GIFT), an open-source Adaptive Instructional System (AIS) architecture that delivers real-time assessments, adaptive coaching, and AARs by synchronizing multimodal learner data (Goldberg et al., 2021; Goldberg & Spain, 2023; Hernandez et al., 2022). GIFT connects to various learning environments through gateway modules and leverages a Domain Knowledge File (DKF), which defines task schemas, competencies, performance indicators and feedback rules. Performance evidence is represented using xAPI (Experience API) statements, a data standard that encodes learning actions, assessments, and contextual information. GIFT's xAPI statements provide information regarding formative (real-time) and summative (overall) assessments aligned to defined tasks and competencies for a given scenario. These xAPI statements are stored in a Learning Record Store (LRS), and can be visualized through performance dashboards, enabling actionable insights for instructors and unit leaders. GIFT includes several modules associated with traditional intelligent tutoring systems including a domain and task module that hosts the DKF and facilitates task assessments, a learner module that gathers information regarding learner states, and a pedagogical and coaching module that controls feedback and remediation strategies (Figure 1). GIFT also includes the Game Master interface that facilitates a "human in the loop" approach to assess performance and inject scenario adaptations during collective simulation-based training events.

To support structured competency modeling, STEEL-R incorporates the Competency and Skills System (CaSS), CaSS manages digital competency frameworks and employs longitudinal learner modeling techniques to determine a learner's competency state in relation to a given framework. These modeling techniques utilize xAPI data and the associated metadata captured within to create competency assertions that serve as inputs into a probabilistic proficiency math model that establishes competency states based on all evidence captured. Together, these components form a data-driven training ecosystem and demonstrable instantiation of the TLA aligned to experiential learning best practices and modeling assumptions.

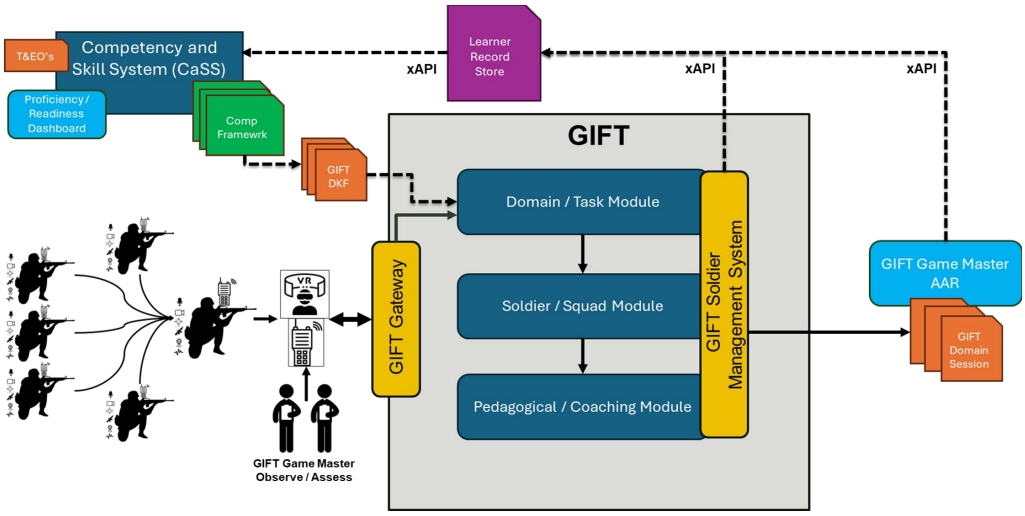

Figure 1. GIFT and CaSS in the STEEL-R Data Architecture

STEEL-R Data Flow, Functionality, and Assumptions

The STEEL-R architecture supports both real-time and post-exercise data workflows. During a training event, data captured from sensors (e.g., GPS, IMU, audio), instrumentation systems (e.g., MILES), and OC/T (e.g., OC/T assessments) are synchronized and processed to produce performance metrics aligned to predefined task steps associated with a task or battle drill in GIFT's DKF. Key features of the STEEL-R data architecture include:

- Real-time ingestion of data streams
- Timestamped synchronization of multimodal data, aligned to scenario events and tasks
- Mapping of raw data to task-relevant behavioral indicators and corresponding tasks and competencies
- Generation of xAPI statements for both formative and summative assessment
- Visualization of performance through custom AAR dashboards
- Persistent tracking of learner states within a competency model

These capabilities allow STEEL-R to provide instructors, OC/Ts, and unit leaders with actionable feedback that is both timely and contextually grounded.

The STEEL-R architecture is built on several core principles that guide its design and implementation. First, it relies on a well-defined competency framework that links core tasks and subtasks to the knowledge, skills, and abilities required for successful performance, along with corresponding performance metrics. Second, it assumes the inclusion of training experiences that provide opportunities to measure task performance. In the case study described next, BD2A served as the task for these training experiences. Third, it incorporates a measurement model that maps tasks and subtasks to observable behaviors and links them to performance metrics and metadata, generating credible evidence of competence. Finally, the architecture uses data standards that are compatible with the Total Learning Architecture (TLA), such as IEEE's Experience API (xAPI), to gather, integrate, and save performance data across training experiences.

In our case study, these principles were operationalized through four key steps: (1) establishing a competency framework for BD2A; (2) using BD2A (React to Contact) as a live training experience, providing multiple opportunities to assess team behaviors—including formation maintenance, fire coordination, and communication—using data captured from GPS, IMU, and audio sensors; (3) instantiating the competency framework into an actionable measurement model (e.g., domain knowledge file) within GIFT that links multimodal data with metrics of performance, and (4) generating and passing xAPI data to CaSS to stimulate competency assertions.

CASE STUDY: EXTENDING STEEL-R TO LIVE TRAINING ENVIRONMENTS

While STEEL-R has been validated using simulated data (Robson et al., 2022), further real-world testing is needed to mature its competency modeling methods and to assess its ability to capture, process, and analyze disparate data sources for timely AAR sessions under live training conditions. To address this need, we tested STEEL-R during a series of live situational training exercises (STX) that involved four squads (n = 4) completing iterations of BD2A over the course of two weeks. Each squad completed 4 iterations of the STX lane. Squad members were equipped with multiple sensors, including GPS units, inertial measurement units (IMUs) mounted on helmets and weapons, and microphones to capture individual location data, movement and weapon orientation, and team communications. Squad members used the Multiple Integrated Laser Engagement System (MILES) Integrated weapon kit, a laser-based, force-on-force, combat simulation system, to stimulate and log simulated engagements and hits with opposing force (OPFOR) elements.

To begin the exercise, squads were given operational orders to advance toward a designated objective while searching for enemy combatants. OPFOR engaged the squads, prompting them to respond with appropriate tactics, and ultimately complete an assault on the objective. Successful performance required squad members to coordinate actions, exchange information, and adapt to changing circumstances. Each battle drill took 20-30 minutes to complete. Trained OC/Ts (n = 3) rated each squad's performance using a three-point scale of "below expectation", "at expectation", and "above expectation." These ratings were made for each task associated with the battle drill using GIFT Game Master which ran on a handheld tablet. At the completion of each battle drill, squad members completed post-exercise self-report surveys to assess teamwork, communication, and performance outcomes and participated in an AAR led by the OC/Ts. Squads were randomly assigned to either receive an enhanced AAR that included displaying and reviewing the exercise in GIFT Game Master with multimodal data performance visualizations (n = 2) or a non-enhanced AAR (n = 2). The combination of sensor streams and OC/T ratings created a rich, multimodal dataset aligned with our objective to extend STEEL-R and measure small unit performance under the conditions of live training.

In the following sections, we describe the architectural enhancements we applied to STEEL-R to accommodate multimodal data streams to support more timely and enhanced assessments of squad performance and the competency model framework we developed that transforms raw data into actionable insights of task and team performance.

Establishing a Competency-based Assessment Framework

Prior to collecting data, we developed a competency model for BD2A. We extended our previous research using the Hierarchical Affect, Behavior, and Cognition (H-ABC) model of teamwork (Vatral et al., 2022; Vatral et al., 2023), to link domain-specific performance indicators derived from sensor to tasks and high-level teamwork constructs (Figure 2). The H-ABC framework is grounded in evidence-centered design, which infers higher-level competencies from task-level evidence (Mislevy et al., 2003).

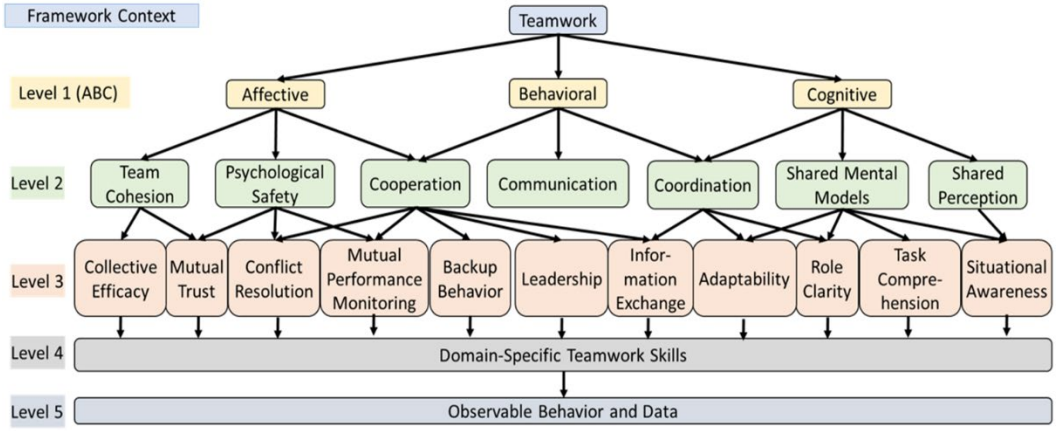

Figure 2. Hierarchical ABC Competency Framework

The first step in developing the model was listing and defining the task steps in BD2A as documented in the T&EO checklist. Using this structure, we framed BD2A in terms of phases and tasks. For each task, we defined observable Soldier/squad behaviors (e.g., "The squad maintains proper formation and spacing between Soldiers while approaching contact") and identified corresponding behavioral measures for assessing that task step (e.g., "Fire team minimum separation distance"). We leveraged metrics, data sources, and thresholds established from the Small Unit Performance Analytics (SUPRA) program, a prior DEVCOM Soldier Center project focused on measuring small unit performance (O'Donovan et al., 2023). The SUPRA program identified predictors of small unit performance derived from sensor data that could be mapped to performance constructs—such as movement, fire team effectiveness, and coordination. We leveraged this research and the corresponding metrics to map task-steps from BD2A with observable behaviors and data sources that could serve as evidence of performance. These measures were then linked to overarching teamwork competencies within the H-ABC framework. For instance, GPS data provided input for the tasks of "Fire team maintains minimum separation distance" which was mapped to the team function of "Role Clarity", illustrating how performance metrics can serve as evidence of more complex team functions. This structured mapping—from raw sensor data to task behaviors to team competencies—served as the foundation for automated, competency-based assessments within STEEL-R.

Because STEEL-R's architecture supports real-time, continuous data streaming, accurate performance scoring requires precise task identification within the drill. To ensure that our competency model and measurement framework analyzed only relevant data, we segmented the competency model into five phases. This allowed our model to isolate data specific to each task execution timeframe. For example, evaluating the "Approach" phase of the drill focused solely on GPS data from that period, excluding data from other phases.

In addition, we included tentative performance standards and thresholds for each of the behavioral measures in the model. We did this so the measures could provide preliminary performance standards aligned to performance thresholds for "below expectation", "at expectation", and "above expectation". We drew on authoritative sources such as field manuals and data gathered from SUPRA to inform these thresholds. For instance, the *Ranger Handbook* recommends a minimum ten-meter separation between Soldiers in open terrain (Department of the Army, 201), so a "below expectation" rating for the "Fire team minimum separation distance" measure would apply when spacing falls short of that threshold. Although these standards are provisional, STEEL-R's data pipeline is intended to support ongoing efforts to refine and formalize the performance norms in future applications.

STEEL-R Enhancements for Live Training Integration

Along with developing the competency model, we implemented a series of enhancements across STEEL-R and GIFT, expanding their capabilities to support real-time, multimodal assessment in live training environments. These enhancements fall into five primary areas: sensor integration, data synchronization and alignment, data integration and playback, communication analysis, and AAR visualization and scoring (Table 1)

Sensor Integration
We expanded STEEL-R's data ingest pipeline to incorporate a wider range of sensor data, including MILES data for capturing simulated weapon engagements, GPS data for gathering positional and movement data, IMU sensors for orientation data, and audio recordings from body-worn microphones. Furthermore, we updated STEEL-R to operate with a mesh network established using tactical radios, enabling real-time positional and movement analytics.

Data Synchronization & Alignment
To enable real-time assessment, we focused on improving data synchronization and alignment. We developed a new interface within GIFT that enables efficient synchronization of time-stamped data streams—such as IMU, audio, and MILES—based on squad roles and positional identifiers. This interface allows individual-level data (e.g., from the Alpha and Bravo Team Leaders) to be accurately associated with corresponding tasks and behaviors, while also aggregating these inputs to produce a synchronized, team-level performance view. By structuring data alignment around squad hierarchy and temporal context, GIFT now supports both granular and collective analyses that feed directly into competency modeling and AAR visualizations.

Data Integration and Playback
To provide synchronized playback and review of multimodal performance data, we refined GIFT to ingest compiled domain session files containing GPS, audio, MILES, IMU, and OC rating data. Once ingested, GIFT aligns these data

streams temporally and replays the session at 5x speed to support rapid review. The result is a "review session" file that can be launched within the GIFT Game Master interface, providing a game-time visualization of squad performance across mission phases and tasks. This output includes synchronized displays of Soldier locations on a map, IMU-based heading and movement data, communication transcripts, and task-level BD2A ratings—all structured to guide data-informed AARs.

Communication Analysis
To support efficient integration and processing of audio data, we extended GIFT's external assessment engine to work with a voice-to-text transcription service which provides transcripts of audio recording gathered during training sessions. GIFT manages the compilation of the audio files and gathers the transcripts, which are then synchronized with the playback session, allowing communications data to be processed and visualized alongside sensor metrics.

AAR Visualization and Scoring
To better integrate Observer/Trainer (OC/T) driven assessments, we modified GIFT to support OC/T session-sharing behavior, allowing automated and OC-driven assessments from multiple OCs to be aggregated into a single unified assessment session. This enabled OCs to easily share their session observations with one another to collaboratively review, discuss, and reconcile ratings before finalizing consensus ratings for each task. These upgrades were paired with a redesigned GIFT Game Master interface that streamlined the Observe–Assess–Review workflow, allowing OCs to easily take notes and capture voice memos during a drill. GIFT time-aligns these notes in Game Master, enabling OCs to easily preview them during an AAR playback session.

Table 1. Enhancements to STEEL- R and GIFT Functional Areas

Functional Area	Enhancement Description	System Component(s)
Sensor Integration	GPS and IMU streaming via Athena-Tek radios and TAK Server	STEEL-R
	MILES integration for weapon engagement logging	STEEL-R
	Body-worn microphones for audio capture	STEEL-R / GIFT
Data Synchronization & Transmission	Real-time synchronization of multimodal data	STEEL-R
	Stand-alone deployment via Android VM with Linux SDK improvements	STEEL-R
Data Integration and Playback	Extended GIFT DKF for hierarchical competency modeling	GIFT
	SUPRA-derived condition classes (e.g., Movement, Security)	GIFT / STEEL-R
	Updated xAPI profile for contextualized, time-aligned assessments and environment condition metadata tagging	STEEL-R
	OC session sharing: aggregation of OC-driven and automated assessments across GIFT instances	GIFT
Communication Analysis	Voice-to-text transcription via TDMS for communications analysis	GIFT
AAR Visualization & Scoring	Redesigned Game Master UI for Observe–Assess–Review workflows and enhanced top-down map view	GIFT Game Master / STEEL-R Dashboard
	Real-time scoring and performance summaries	STEEL-R

In addition, we updated the Competency and Skills System (CaSS) with a new hierarchical competency framework supporting H-ABC framework, establishing a parental relationship between the H-ABC framework and the domain-specific subtasks required by BD2A. Both frameworks were imported into GIFT to instantiate the DKF that was used to evaluate sensor metrics gathered during an exercise, ensuring that measures of team performance represented in xAPI were linked back to their originating competency definitions in CaSS. Incorporating these frameworks shaped

several changes made to GIFT's DKF schema, namely, to ensure that multiparent relationships between competencies were preserved in the DKF model. CaSS was further equipped with Bayesian Knowledge Tracing and Generalizability Theory models to estimate competency states and assessment reliability using the xAPI statements produced from GIFT using the DKF made from these frameworks.

These enhancements collectively enabled STEEL-R to capture, synchronize, and process the multimodal data necessary to provide timely and actionable feedback during live training exercises.

CASE STUDY RESULTS AND KEY OBSERVATIONS

Data Processing

A primary goal of our case study was to evaluate whether the enhancements we implemented to STEEL-R supported the ability to collect, synchronize, and process multimodal data in near real time to support timely assessments and AARs. Testing involved instrumenting Soldiers with sensors to gather and process data from squads after each STX lane iteration. During week 1, processing time for each battle drill iteration – encompassing gathering, collecting, uploading, and parsing GPS, IMU, MILES, audio recordings, and observer ratings averaged approximately 90 minutes between uploading the data and being able to process and visualize the data in Game Master to support an AAR. This exceeded the threshold for timely AAR delivery and highlighted the need for workflow and system optimization.

Through iterative improvements to the STEEL-R data pipeline, including parallelized data ingestion, real-time synchronization, and automated session configuration, we substantially reduced this turnaround time. By week two, a complete iteration could be processed in approximately 35 minutes, meeting the operational requirement for timely AARs and confirming STEEL-R's feasibility for deployment in dynamic field environments where rapid feedback is critical. Despite these improvements, several technical challenges emerged during implementation. Specifically, we encountered GPS drift and unreliable streaming of IMU and MILES data over the mesh network. These limitations necessitated manual extraction and uploading of affected data streams, introducing delays, and increasing the overall latency of the processing pipeline.

Enhancing AAR Engagement through Game Master-Driven Data Visualizations

Another key goal of the use case was examining whether performance visualizations facilitate through STEEL-R enhanced reflective learning and the perceived effectiveness of AARs. To investigate this, we analyzed self-report data from participating Soldiers who completed a 5-item survey immediately following each AAR session. The survey assessed how effectively the AAR helped Soldiers understand and improve individual and team performance across several performance dimensions. Responses were collected using a 5-point Likert-type scale ranging from 1 (ineffective) to 5 (completely effective). Example items included "How effective was the AAR at providing feedback on how to improve my squad's movement?" and "How effective was the AAR at providing feedback on how to improve my squad's coordination?"

Analysis of the survey responses showed that data visualizations presented during the AAR using multimodal data—such as synchronized GPS trajectories, communication transcripts, and IMU, and MILES-based engagement data — were highly effective at improving squad movement, communication, and coordination over the course of the case study. Mean ratings were uniformly high for all three dimensions (coordination, squad movement, and communication), reflecting strong perceived utility of the data-enhanced AAR format (Table 2). Repeated measures ANOVAs revealed no statistically significant differences in perceived effectiveness for coordination from Session 1 ($M = 5.00$; $SD = 0.00$) to Session 4 ($M = 4.86$; $SD = 0.54$), $F(3, 39) = 1.00$, $p = .43$. Similarly, no significant differences were found across sessions for movement, $F(3, 39) = 1.00$, $p = .40$, or communication, $F(3, 39) = 1.64$, $p = .10$.

Comparisons between squads that received AARs with multimodal data visualizations and those that did not failed to reach statistical significance for movement ($F(1, 27) = 1.31$, $p = .26$), communication ($F(1, 27) = 0.28$, $p = .87$), and coordination ($F(1, 27) = 0.35$, $p = .56$), likely due to the limited sample size available for between-group analysis.

These results suggest that participants found the data-driven AARs informative, actionable, and beneficial for supporting team reflection. The feedback obtained through informal discussions reinforced that the inclusion of synchronized GPS trajectories, MILES engagement outcomes, time aligned communication transcripts, and IMU-

based movement maps helped contextualize feedback and stimulated more focused and productive discussion among squad members.

Table 2. Average Effectiveness Ratings of Data Enhanced and Non-enhanced AARs on Performance Competencies

Session	Movement		Communication		Coordination	
	No Visualization M (SD)	Data-Visualization M (SD)	No Visualization M (SD)	Data-Visualization M (SD)	No Visualization M (SD)	Data-Visualization M (SD)
1	4.73 (0.70)	5.00 (0.00)	4.80 (0.78)	5.00 (0.00)	4.71 (0.73)	5.00 (0.00)
2	5.00 (0.00)	4.86 (0.54)	4.80 (0.78)	4.71 (0.73)	5.00 (0.00)	5.00 (0.00)
3	4.87 (0.52)	5.00 (0.00)	5.00 (0.00)	5.00 (0.00)	5.00 (0.00)	5.00 (0.00)
4	4.80 (0.78)	5.00 (0.00)	4.87 (0.52)	4.86 (0.54)	5.00 (0.00)	4.86 (0.54)

DISCUSSION AND FUTURE WORK

The goal of our case study was to test whether STEEL-R could ingest- process and transform these data into task-aligned assessments and data visualizations to support timely AARs. The results of our case study demonstrate the feasibility and operational utility of integrating multimodal assessment capabilities into live infantry training using STEEL-R. Through targeted enhancements to STEEL-R's architecture and GIFT, we were able to gather, synchronize, and process multimodal data to deliver actionable assessments and enhanced AAR experiences. Notably, we were able to reduce data processing time for each battle drill iteration from approximately 90 minutes to approximately 35 minutes over the course of our case study, enabling timely feedback delivery during AARs. Furthermore, Soldiers reported consistently high ratings of AAR effectiveness, particularly in supporting squad movement, communication, and coordination across battle drill iterations, underscoring the instructional value of leveraging multimodal data to visualize and augment AARs and support team learning and performance improvement.

These findings have important implications for the design of adaptive training systems and the use of multimodal learning analytics in military contexts. By extending the STEEL-R architecture to live training we have moved towards capturing timely and objective performance feedback to support data-driven assessments and structured experiential reflection. These enhancements support the key principles of experiential learning—particularly the role of meaningful reflection in correcting misconceptions and supporting skill development. Future research should continue to extend and apply STEEL-R's data strategy and framework to support additional training contexts. While our case study demonstrated the feasibility of integrating multimodal data into STEEL-R during a small-scale training event, further work is needed to evaluate and optimize its data processing workflow under more demanding, large-scale training exercises. During the case study, we encountered several technical challenges, including GPS drift and the inability to reliably stream IMU data and MILES data over our network. As a result, we were required to manually extract and upload these data sources into STEEL-R, increasing the latency of the processing pipeline. Addressing these limitations will require additional architectural enhancements to improve the speed and accuracy of multimodal data alignment and integration. More seamless integration of sensor data, including GPS, IMU, and MILES systems—over a mesh network will also be essential for achieving near real-time performance assessments and immediate AAR capabilities.

Another area for future development involves expanding the use of communication data within the assessment process. While STEEL-R can produce and integrate transcripts of voice communications, leveraging natural language processing (NLP) techniques to enable automated assessment of key communication competencies offers exciting possibilities to assess factors mediating unit performance. Furthermore, future research should focus on scaling the STEEL-R architecture to a broader range of training scenarios and tasks. Extending the GIFT and the competency model to cover additional battle drills and mission sets could enable more widespread adoption of automated assessment tools across units, competencies, and data types.

CONCLUSION

Advancements in virtual and live training technologies are providing new opportunities to leverage multimodal data—video, audio, sensor, geographical positioning, and behavioral data—to deliver objective performance assessment. Reliably capturing, synchronizing, and interpreting these multimodal sources is critical for identifying performance gaps and assessing training effectiveness. This paper discusses how we have systematically extended STEEL-R's architecture to reliably capture, synchronize, and interpret multimodal sources and how we implemented an evidence-centered framework for converting raw data into actionable insights that can support performance assessment. Multimodal learning analytics offers a promising approach for supplementing expert assessments and feedback. Continuing to investigate and develop tools and methods that can support multimodal data-informed insights for squad performance is imperative for advancing next generation training capabilities and enhancing Soldier readiness.

REFERENCES

Acosta, H., Lee, S., Mott, B., Bae, H., Glazewski, K., Hmelo-Silver, C., & Lester, J. (2024). multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning. P*roceedings of the Seventeenth International Conference on Educational Data Mining*, pp. 224-235, Atlanta, Georgia, 2024.

Ainam, J. P., Yanik, E., Rahul, R., Kunkes, T., Cavuoto, L., Clemency, B., ... & De, S. (2025). Deep learning for video-based assessment of endotracheal intubation skills. *Communications Medicine, 5*(1), 116.

Azevedo, R., Dever, D., Wiedbusch, M., Brosnihan, A., Delgado, T., Marano, C., ... & Smith, K. (2024, October). A taxonomy for enhancing metacognitive adaptivity and personalization in serious games using multimodal trace data. In *Joint International Conference on Serious Games* (pp. 27-40). Cham: Springer Nature Switzerland.

Blake-Plock, S., Owens, K., & Goodell, J. (2023, July). The value proposition of GIFT for the field of learning engineering. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)* (p. 87-94). US Army Combat Capabilities Development Command–Soldier Center.

Emerson, A., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2023). Multimodal predictive student modeling with multi-task transfer learning. P*roceedings of the Thirteenth International Learning Analytics and Knowledge Conference*, pp. 333-344, Arlington, TX.

Giannakos, M., & Cukurova, M. (2023). The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology*, *54*(5), 1246-1267.

Goldberg, B., Owens, K., Gupton, K., Hellman, K., Robson, R., ... & Hoffman, M. (2021). Forging competency and proficiency through the synthetic training environment with an experiential learning for readiness strategy. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL*.

Goldberg, B., & Spain, R. (2023, July). Supporting Future Learning Concepts: GIFT in the Year 2040. In *Proceedings of the 11th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym11)* (p. 19-25). US Army Combat Capabilities Development Command–Soldier Center.

Hancock, C. L., Teo, G. W., King, M. J., Goodwin, G. A.,... & O'Donovan, M. P. (2025). Quantitative team performance metrics for dismounted infantry battle drill analysis. *Applied Ergonomics*, *125*, 104473.

Hernandez, M., Blake-Plock, S., Owens, K., Goldberg, B., Robson, R., Center, S., & Ray, F. (2022). Enhancing the total learning architecture for experiential learning. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL*.

Johnson, C., & Gonzalez, A. J. (2008). Automated after action review: State-of-the-art review and trends. *The Journal of Defense Modeling and Simulation*, *5*(2), 108-121.

Johnston, J. H., Sinatra, A. M., & Swiecki, Z. (2020). Application of team training principles to visualizations for after-action reviews. In *Design Recommendations for Intelligent Tutoring Systems: Volume 8 - Data Visualization*. (Sinatra, A.M., Graesser, A.C., Hu, X., Goldberg, B., and Hampton, A.J. (Eds.), pp 19-29, Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center.

Keiser, N. L., & Arthur Jr, W. (2022). A Meta-Analysis of Task and Training Characteristics that Contribute to or Attenuate the Effectiveness of the After-Action Review (or Debrief). *Journal of Business and Psychology*, *37*(5), 953-976.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Englewood Cliffs, NJ: Prentice-Hall.

Kolb, D. & Kolb, A. (2017). *The experiential educator: Principles and practices of experiential learning*. EBLS Press. Kaunakakai, HI 96748.

Meliza, L. L., & Goldberg, S. L. (2017). Impact of after-action review on learning in simulation-based US Army training. In *Assessment of problem solving using simulations* (pp. 255-272). Routledge.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military medicine*, *178*, 107-114.

Morrison, J. E., & Meliza, L. L. (1999). Foundations of the after action review process (Special Report 42). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

O'Donovan, M. P., Hancock, C. L., Coyne, M. E., Racicot, K., & Goodwin, G. A. (2023). Assessing the impact of dismounted infantry small unit proficiency on quantitative measures of collective military performance Part 1: Recommended test methodologies. Technical Report TR-23/013. US Army Combat Capabilities Development Command–Soldier Center.

Owens, K., & Goldberg, B. (2022). Competency-based experiential-expertise. In *Design Recommendations for Intelligent Tutoring Systems*, Volume *9*, Competency-based Scenario design (Eds A. Sinatra, A., Graesser, X., Hu., B., Goldberg, A., Hampton, & JH. Johnston) pp. 19-29. US Army Combat Capabilities Development Command–Soldier Center.

Owens, K., Townsend, L., Goldberg, B., Abrams, J., and Cooke, G. (2024). Learning engineering competency-based experiential learning within military institutional training and education. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

Robson, R., Ray, F., Hernandez, M., Blake-Plock, S., Casey, C., Hoyt, W., ... & Goldberg, B. (2022). Mining artificially generated data to estimate competency. *Proceedings of the 15th International Educational Data Mining Society*. Retrieved from https://educationaldatamining.org/edm2022/proceedings/2022.EDM-industry-track.110/index.html.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human factors*, *50*(6), 903-933.

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

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, *51*(5), 1450-1484.

Smith, A., Spain, R., Min, W., Goldberg, B. S., & Lester, J. C. (2023). Towards a Multimodal Data-driven Framework for Adaptive Coaching in Collective Simulation-Based Training. In AI-GEL@ AIED (pp. 40-47).

U.S. Army Training and Doctrine Command. (2024). The Army Learning Concept for 20302040. TRADOC.

US Department of the Army (2019) Army Modernization Strategy: Investing in the Future (Washington, DC: Government Printing Office, 2019),

Vatral, C., Biswas, G., Mohammed, N., & Goldberg, B. S. (2022). Automated assessment of team performance using multimodal Bayesian learning analytics. . Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

Vatral, C., Biswas, G., Mohammed, N., and Goldberg, B. S. (2023). A framework for performance assessment across multiple training scenarios using hierarchical Bayesian competency models. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.