Behavior-Based Performance Optimization in Emerging Training Environments

Audrey Zlatkin, Ph.D., Costas Koufogazos, Gwen Campbell, Ph.D., Peyton Bailey, William Rivera, Ph.D.
Design Interactive, Inc
Orlando, FL
audrey.zlatkin@designinteractive.net, costas.koufogazos@designinteractive.net,
gwen.campbell@designinteractive.net, peyton.bailey@designinteractive.net,
william.rivera@designinteractive.net

ABSTRACT

The DoD is strategically pursuing advanced technology solutions to enhance existing training systems and prepare its warfighters for high-end combat. Traditional means of assessing student performance are subjective, generalized, and subject to bias, presenting challenges when adapting to emerging training environments. There is a need to accelerate skill acquisition in any context through adaptive and personalized training, standardized assessments, and contextualized performance feedback/insights. Additionally, leveraging Artificial Intelligence (AI) for data analytics and performance modeling has the potential to reduce subjectivity for military training, accelerating learning and optimizing performance outcomes across the DoD. Design Interactive, supported by funding from the Office of Naval Research and Air Force Global Strike Command, applied a research-based approach to develop a sustainable and extensible training and evaluation tool that provides assessment of performance through blending instructor insights with standardized, competency-based rubrics, and structured After-Action Review (AAR).

To meet the needs of modernizing training environments, a method for integrating simulator data to provide automated, AI-driven performance insights and analytics was conceptualized. Outcomes of this effort have resulted in an approach for integrating virtual flight training data to provide standardized, human-in-the-loop performance assessment combined with automated performance metrics, reducing subjectivity and accelerating learning across military training environments. This paper discusses the behavioral assessments that are foundational to this system, the development of a performance algorithm to train the system's AI from real-world user data, and results from user testing that continue to inform designs for more contextualized feedback and additional assessment methods to be incorporated into future versions of the software. Design Interactive aims to enhance warfighter competence, reduce training costs, and improve mission readiness. This work is continuing to inform R&D efforts across the Marine Corps and Air Force and has been tested within a wide range of use cases from military, medical, academic, and industrial settings.

ABOUT THE AUTHORS

Dr. Audrey Zlatkin is a Senior Research Associate at Design Interactive, Inc. (DI) with over 9 years of experience in human factors, cognitive psychology, and human-systems integration with a focus on adaptive training and decision support technology for operational environments.

Costas Koufogazos is a Research Associate III at Design Interactive, Inc. (DI) with 6 years of expertise in Human Factors and research. His current projects consist of efforts to design and develop tools used by the military for behavior-based performance optimization, used to accelerate skill acquisitions across different contexts.

Dr. Gwen Campbell is a Senior Research Associate III at Design Interactive, Inc. (DI) with over 19 years of experience in human performance and human systems integration, simulation, applied statistics, operations research, lean six sigma, and project management.

Peyton Bailey is a Research Associate II at Design Interactive, Inc. (DI) and has over 6 years of experience human factors research and design, training effectiveness evaluations, and developing virtual/augmented reality training solutions for military and medical training.

Dr. William Rivera is Software Systems Engineer IV at Design Interactive, Inc. with more than 20 years of expertise in Data Science, Research, Software Development and Project Management.

Behavior-Based Performance Optimization in Emerging Training Environments

Audrey Zlatkin, Ph.D., Costas Koufogazos, Gwen Campbell, Ph.D., Peyton Bailey, William Rivera, Ph.D.
Design Interactive, Inc
Orlando, FL
audrey.zlatkin@designinteractive.net, costas.koufogazos@designinteractive.net, gwen.campbell@designinteractive.net, peyton.bailey@designinteractive.net, william.rivera@designinteractive.net

INTRODUCTION AND BACKGROUND TO THE PROBLEM STATEMENT

In modern military strategy, the enhancement of training systems through advanced technology has become a pivotal focus. As both adversaries and technology evolve, it is essential that airmen are prepared for high-end combat environments. Over the past decade, there has been a significant rise in virtual training systems. Innovations in Extended Reality (XR) and simulator-based solutions are transforming both cognitive and physical skill acquisition, making training more cost-effective and efficient through immersive, repeatable experiences. However, this technological revolution raises critical questions about learning in virtual environments. With diverse training opportunities, how does the Department of Defense (DoD) manage standardized evaluation and performance tracking? Are the investments of time, money, and resources yielding the desired results? While there is an emphasis on developing new training capabilities, it is equally vital to focus on technologies and approaches that ensure consistent, standardized assessment across various aspects of human performance.

Current methods for evaluating training effectiveness, particularly in pilot training, remain largely unchanged, relying on manual processes and subjective instructor assessments. There is a pressing need for an adaptive and personalized training and assessment approach that leverages emerging technologies to provide standardized feedback and better transfer of skills to operational environments. By incorporating cutting-edge innovations such as artificial intelligence and advanced simulation tools, military training programs can deliver more immersive and effective experiences. These advancements not only enhance the readiness and capabilities of warfighters but also ensure they are equipped to handle the complex challenges of modern combat. Continuous development and implementation of advanced training technologies are crucial to maintaining a proficient and adaptable military force.

Challenges In Traditional Assessment Methods

Traditional assessment methods in military training present several significant challenges, primarily due to their subjectivity. These methods often rely heavily on the personal judgment of instructors, which can vary widely. While the empirical knowledge of expert instructors is invaluable, this subjectivity can lead to inconsistent and generalized assessments that may not accurately reflect the individual capabilities and progress of each warfighter (Malone, Vogel-Walcutt, Ross, & Phillips, 2014). Such bias can result in unfair evaluations, impacting the development and confidence of trainees. On the flip side, checklist-driven assessments, though less prone to individual bias, lack the nuanced personalized feedback necessary for comprehensive evaluation. These limitations underscore the need for more objective, robust, and equitable assessment tools that can accurately measure warfighter skills and readiness.

Instructors play a crucial role in the assessment and development of warfighters, but traditional evaluation methods can be a substantial burden on human evaluators. Manually observing, recording, and assessing each trainee's performance is time-consuming and labor-intensive, often detracting from instructors' ability to focus on safety, teaching and mentoring. Instructor pilots are particularly vulnerable to the effects of fatigue due to high workloads and intensive training schedules (McDale & Ma, 2008). Advanced technological solutions, such as automated assessment and analytics tools, can significantly reduce this workload. These technologies offer real-time, objective feedback and detailed performance insights, allowing instructors to allocate more of their time and energy to enhancing the training experience and addressing the individual needs of each warfighter. This shift improves the efficiency and effectiveness of training programs, as well as the quality of instruction and support provided to trainees.

Advantages Of Virtual & Simulation Based Training

Virtual and simulation-based training offers numerous advantages over traditional training methods, especially in military environments. These technologies create highly immersive and realistic scenarios, enabling trainees to experience and respond to situations they are likely to encounter in the field. This approach reduces the risks and costs associated with live training exercises, such as wear and tear on equipment and the need for extensive manpower (Lele, 2013). Moreover, virtual training systems can be tailored to individual learning paces and needs, ensuring that each trainee receives the most effective instruction. Simulations also facilitate repetitive practice, allowing trainees to hone their skills through continuous exposure to diverse scenarios. Additionally, virtual and simulation-based training can be conducted in various locations and at convenient times, increasing training capacity and access to performance metrics (Stanney et al., 2022).

These training systems generate a vast amount of data on trainee performance, usage patterns, and variability across users and scenarios. This data provides insights into individual trainees' strengths and weaknesses and can be leveraged organization-wide to develop more targeted and effective training programs. By integrating virtual and simulation-based technologies, military training not only enhances the realism and efficacy of the training but also improves the preparedness and adaptability of warfighters.

Our research aims to support consistent standardized pilot assessment, reduce instructor bias, and increase training capacity and performance tracking through the integration of research-based learning techniques and AI-driven performance modeling. We propose an approach that integrates automated performance and event data from virtual training systems with human-in-the-loop instructor evaluation. This method provides a comprehensive picture of a trainee's knowledge, skills, and cognitive state within a cohesive after-action review. We expect that an advanced integrated performance dashboard will streamline performance reviews, presenting information in a way that allows students and instructors to quickly identify and address skill gaps.

Research Approach

To enhance training outcomes, we utilized an existing prototype behavior-based training and assessment platform designed to effectively evaluate critical operational behaviors and communications. This platform combines standardized performance evaluations with structured after-action reviews and personalized feedback, expediting skill mastery through observation. By expanding its capabilities, we transformed this platform into a versatile research and data collection tool to explore adaptive, automated training techniques within an initial simulation training use case.

Our approach incorporates various evidence-based training strategies to heighten the efficacy of warfighter training and assessment. Interventions such as behavior-based performance assessment, video self-modeling, self-reflection, and adaptive personalized feedback were selected based on their proven potential to enhance training practices, particularly when integrated with emerging technologies. For example, video self-modeling (VSM) has shown significant potential to accelerate skill acquisition and optimize performance, making it especially beneficial in military training contexts (Dowrick, 1999). Recent research by Yu et al. (2020) demonstrated the value of video modeling and feedback in procedural motor skill acquisition, showing a significant reduction in time to mastery for medical skill training. This approach leverages observational learning and self-reflection, allowing individuals to internalize mistakes and correct them in future practice scenarios.

Personalized feedback is crucial in our training intervention, offering targeted, individualized guidance tailored to each trainee's specific performance gaps and learning requirements. This approach pinpoints areas for improvement, reinforces positive behaviors, and accelerates skill development (Hattie & Timperley, 2007). Additionally, research by Shute (2008) indicates that targeted and elaborative feedback, which points out specific behaviors and responses beyond just ‘correct/incorrect,’ has a significantly greater impact on learning outcomes.

In the current work, we explored enhancing the training and feedback process by leveraging machine learning (ML) techniques for adaptive feedback. Incorporating ML techniques allows organizations to extract valuable insights from data patterns, identify skill gaps, predict future performance, and customize training programs to meet specific needs. This data-driven approach uncovers hidden correlations that may elude traditional analysis methods, thereby empowering informed decision-making and enhancing training efficiencies.

Structured performance evaluation techniques are pivotal in mitigating the subjective biases and inconsistencies inherent in conventional warfighter training methodologies. Unlike subjective approaches, structured assessments provide a systematic framework for evaluating performance against predefined criteria. A behavior-based performance optimization strategy enhances individual behaviors and cultivates skills and habits crucial for real-world scenarios by pinpointing actions that contribute to successful outcomes.

To support a behavior-based assessment approach, we chose Behaviorally Anchored Rating Scales (BARS) for structured performance evaluation. BARS are based on the Dreyfus and Dreyfus (1986) model, which outlines the five stages of mastery development from novice to expert. BARS identifies observable behaviors associated with different levels of cognitive skill, providing a method for evaluating human cognition, decision-making processes, and rationale. Instructors and evaluators are often prone to bias and subjectivity when assessing performance. Using BARS enhances rater objectivity, leading to greater inter-rater and intra-rater reliability. This competency-based skill approach enables rapid and efficient skill development, helping diagnose proficiency levels and identify performance gaps for learners, instructors, and exercise developers. Training that promotes individual trainee awareness and regulation of their own thinking, behaviors, biases, and competencies has been shown to enhance the perception and understanding of critical cues (Ford et al., 1998), the acquisition of higher-order conceptual knowledge (Fiorella & Vogel-Walcutt, 2011), and the development of appropriate mental models and behaviors within a domain (Wu & Looi, 2012).

In addition to BARS, incorporating procedural rating scales and counter-based rating allows for more robust data collection within USAF training environments, which are heavily checklist-driven. This integration provides a dynamic approach to performance assessment, capturing both qualitative and quantitative data for a more comprehensive performance review.

This research aims to leverage these validated training interventions to develop an adaptive training engine. This engine will analyze data from virtual training technologies, incorporate instructor ratings on performance and critical events, and deliver cohesive after-action reviews enhanced through automated, and AI driven performance insights.

METHODOLOGY

Use Case: B52 Air Refueling

The current work focused on B-52 Air Refueling as the primary use case, with the goal of establishing an approach for streamlined after-action review for simulation-based pilot training. Air refueling refers to the process of transferring fuel from the tanker aircraft to the B-52 bomber while both aircraft are in flight, a critical capability for extending operational range and mission duration. This use case was selected due to a recognized gap between the emerging simulation training systems and the existing performance assessment capabilities.

Outcomes of domain and task analyses with subject matter experts (SMEs) resulted in a contextualized rating rubric tailored to B-52 aerial refueling. This process involved pinpointing the key performance areas (KPAs) and observable behaviors essential for the air refueling task and aligning them with the five stages of mastery: novice, advanced, competent, proficient, and expert. The initial Air Refueling rubric underwent multiple iterations with input from three distinct subject matter experts (SMEs) throughout the 12-month project duration. This collaborative process allowed for the refinement and consolidation of criteria, focusing only on the most pertinent key performance areas (KPAs) and behaviors.

Table 1. Aerial Refueling KPAs & Behaviors

KPAs	Behaviors
Awareness of Visual References	Aircraft Alignment Cues Up/Down Aircraft Alignment Cues Left/Right Aircraft Alignment Cues In/Out Awareness of Rate-of-Change in Position
Precision of Operation	Precise Management of Rate-of-Closure from Pre-Contact to Contact Precision of Yolk Inputs Up/Down Precision of Yolk Inputs Left/Right Precision of Throttle Adjustments Timeliness of Yolk Adjustments Timeliness of Throttle Adjustments Recognize Need for Corrective action/disconnect
Workload/ Composure	Composure Under Stress Observed Workload / Frustration

The rubric utilized a five-point rating scale for each behavior, employing a Likert-type assessment integrated with descriptive behavioral anchors to enhance the standardization and quality of ratings. The scale ranged from 1 (novice performance) to 5 (expert performance), with descriptive anchors delineating the behavior quality at each level of mastery. These descriptive narratives serve as guidelines for observers, assisting them in selecting appropriate ratings based on observed performance. By reducing observer-rater uncertainty and offering clearly distinguishable behavioral indicators, this approach ensures interrater reliability, facilitates data collection, and yields scores that provide trainees with precise insights into areas requiring improvement or reinforcement from instructional performances.

Integrated After-Action Review

Personalized results and feedback play a crucial role in enhancing warfighter performance and promoting continuous improvement. It is imperative that critical instructor feedback is conveyed while still providing consistency and structure through standardized assessment. This approach can be enhanced through the integration of objective, timestamped data from virtual training systems to provide a clear picture of a trainee's knowledge and cognitive state. Leveraging this wealth of data provides an additional layer of performance review, enabling accelerated learning and automated performance insights through comprehensive performance review

Our approach provides analysis and presentation of results to raters and trainees for a standardized and effective after-action review across three levels of performance review: summary, detail, and video-AAR. The data is presented at the summary level to provide a snapshot of overall performance and highlight individual performance gaps. Detail Results support a user in identifying problem areas in individual or team performance. Graphs and charts support comparison of performance across key areas and low-level behaviors, as well as facilitate review of performance over the course of the entire session, enabling a user to pinpoint exactly when those performance breakdowns occur. In the video-based after-action review, the system provides direct links from the data to the video of performance so that raters and trainees can observe what was happening in the training session when ratings were made. This standardized after-action review enhances the feedback process by supporting users as they visualize and identify performance gaps.

The data is first presented at the summary level to provide insight into opportunities for greatest improvement while also reflecting on performance successes. The overall average score for the entire session is shown along with averages by KPA and behavior. Results also include performance-driven prompts, tips, and/or recommendations for additional or future instruction to help the observer guide the trainee toward improvement on their lowest-rated behaviors. The automated flight data was implemented at the summary result level in the form of a Check Ride Summary which provides an at-a-glance overview of performance across three critical performance criteria, Total Contact Duration,

Number of Disconnects, and Longest Contact Duration. While these metrics alone do not give the whole picture of performance, they provide a snapshot to quickly address problem areas for the learner based on automated data analysis. Table 2 outlines the displayed criteria and their measures of success.

Table 2. Check Ride Criteria

Variable	Evaluation Criteria
Contact Duration	Pass: Greater than 5 mins total
Number of Disconnects	Pass: Three or less disconnects
Longest Contact Durations	Personal Best in Mins/Sec

The Check Ride Summary provides an overview of whether the student passed or failed including progress indicators to show how close they were to achieving the 5 mins minimum of contact time, whether they were over or under disconnect fail criteria, and their personal best contact duration to aid them in improving performance in subsequent practice sessions.

The detailed view contains a more in-depth breakdown of average ratings at the KPA and behavior level. When there is flight data associated with a session, you can track automated event flags for contacts and disconnects (with duration) indicated on the behaviors over time graph. This allows the user to make connections between key air refueling behaviors and critical events from the flight scenario. Additionally, users can hover anywhere over the timeline to view more information such as when a rating was made, what the behavior and associated score was, etc.

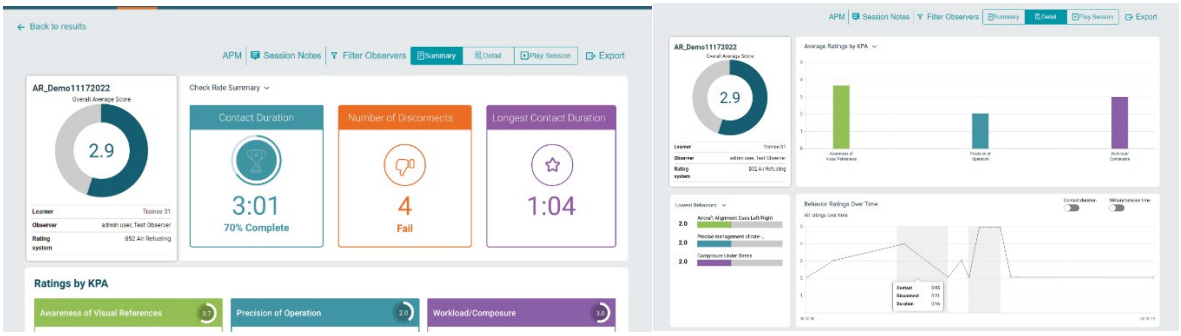

Figure 1: APM Data in Check Ride Summary and Detail

APM data is also viewable in the video-based after-action review (AAR) panel (play session) as shown here. This includes viewing areas of interest (e.g., times of contact) indicated on the video timeline, as well as timestamped event flags for contacts and disconnects from the virtual trainer on the righthand panel along with behavior ratings and associated comments. To support ease of comprehension and intuitive navigation, different iconography was used to differentiate between ratings/comments made by users (i.e., displayed as a round icon with the user's initials) and automated data from the virtual trainer (indicated with a gear icon with an 'A'). Ratings made within regions of contact are easily distinguished by being presented within a gray box (beginning at the contact and ending at the disconnect).

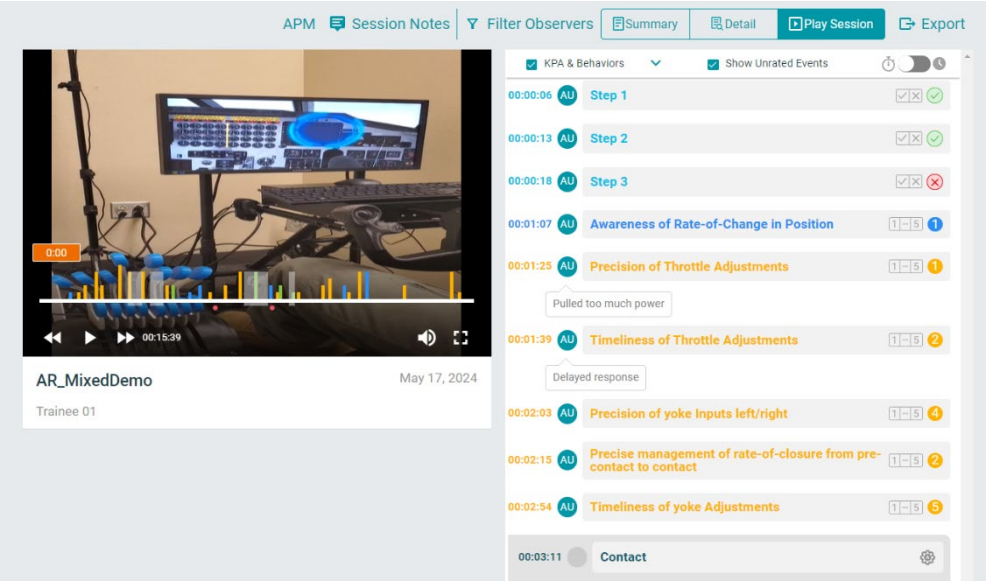

Figure 2: APM in AAR

The APM panel is where raw virtual flight data can be imported and viewed at a summary level. Once data is imported, flight summary metrics (including Automated Performance Insights) are displayed in the top left and timestamped data is displayed on the flight timeline (Figure 5, Left). Variables presented in this panel were identified and refined with stakeholders and SMEs to provide the most value in an overview capacity. Variables include session start time, number of contacts, number of disconnects, time of and to first contact, average contact duration, and overall total contact duration for that flight session. A visuospatial 3D map supports faster comprehension and localization of disconnects (Figure 5, Right). The 3D model provides visual reference of both the boom and aircraft nose to aid better spatialization. This 3D model is also paired with a legend indicating the mechanisms for manipulating the model (zooming, rotating, etc.) as well as short cut buttons that allow the user to snap to the 2D side and front views.

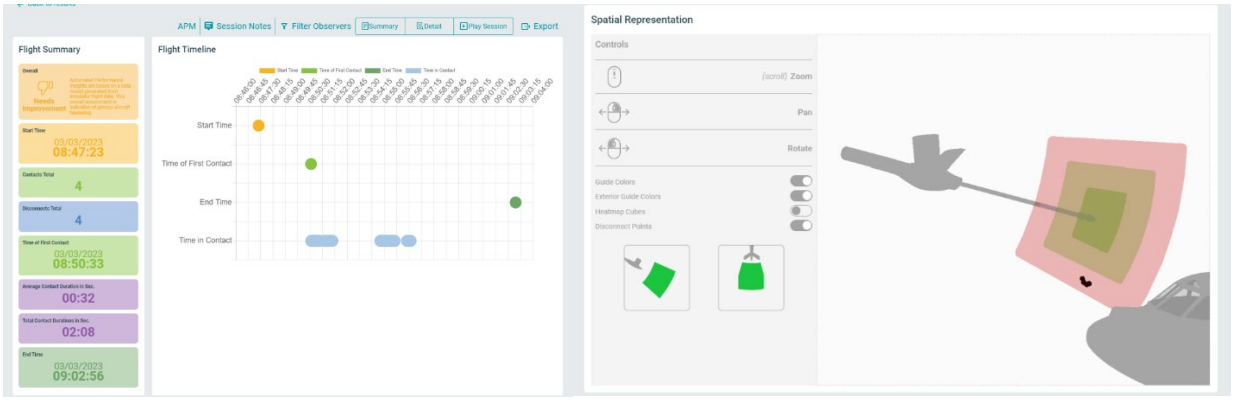

Figure 3: APM Panel

Modeling Approach

Through leveraging AI-driven performance modeling, we have tapped into the potential to provide automated performance insights to trainees, alleviating the workload of human instructors and facilitating large-scale data analytics in future efforts. To achieve this objective, we collected representative data from both expert and novice users. The initial phase involved analyzing 10 samples, evenly split between novices and experts, each accompanied by video recordings of the training session. Novice users were selected from active student pilots while experts were

selected from a pool of Instructor pilots. A SME evaluated these sessions to pinpoint critical behaviors and craft bespoke performance indicators crucial for shaping an adaptive performance algorithm. Utilizing our prototype training and assessment platform, the SME assessed user performance within the simulator, tracking trainer interactions alongside physical inputs on the throttle and yoke. This evaluation process encompassed timestamping specific performance instances and delivering 'goodness' measures for each criterion, utilizing the Key Performance Areas (KPAs) and behaviors outlined in the Air Refueling Rubric.

The behavioral and flight data outputs from both novice and expert performers served as a foundation for constructing our automated performance model. Initial efforts concentrated on a behavior-level, event-based strategy. The primary aim was to rank behaviors and pinpoint qualifying events within the data stream. To ensure data accuracy and relevance, both behavioral ratings and flight data underwent thorough preprocessing. This included compressing flight data into per-second intervals by averaging across milliseconds and synchronizing timestamps for flight events and behavioral ratings within corresponding flight intervals. Additionally, two sets of deltas were computed - the disparity between the tanker and aircraft, and within the aircraft itself - using various time windows to capture nuanced changes.

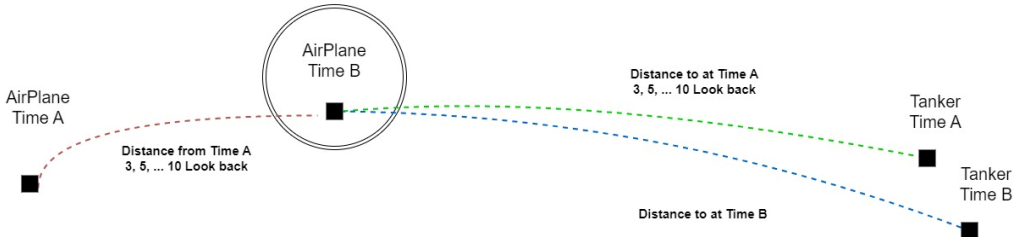

Figure 4. Data Preprocessing

The problem was broken into two fundamental steps. First predict qualifying events and second apply an ML model to determine the ranking. Initially, we devised a ranking model based on behavior deltas over specific time windows, paving the way for a preliminary representation. However, while this approach succeeded in ranking behaviors accurately, identifying qualifying events remained a challenge. This led us to employ multiclass classification to ascertain the presence of qualifying events at any given instance.

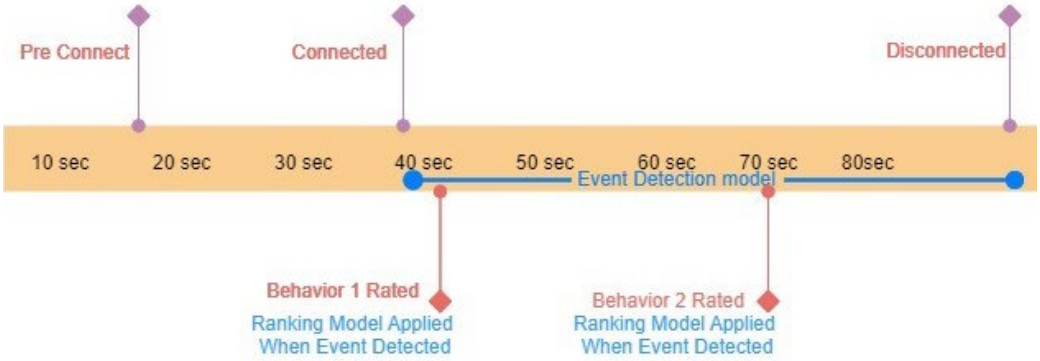

Figure 5. Modeling Approach at the Behavior Level

While several models were applied, Figure 6 represents the ranking model that was ultimately selected as a first pass with an overall accuracy between 72 - 94% that it can correctly classify a behavior plus or minus 1 rank away from the correct rank when in a connected state for the behaviors in Table 2.

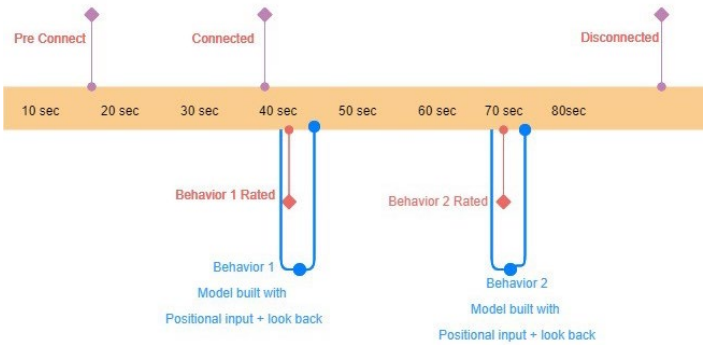

Figure 6. Ranking Model

Table 3: Accuracy of Ranking Model at Behavior Level

Behavior	Accuracy +- 1 rank
Aircraft Alignment Cues In/Out	0.81
Aircraft Alignment Cues Left/Right	1
Aircraft Alignment Cues Up/Down	1
Precision of Throttle Adjustments	0.77
Precision of Yoke Inputs left/right	0.85
Precision of Yoke Inputs up/down	0.9
Timeliness of Throttle Adjustments	0.74
Timeliness of Yoke Adjustments	0.94
OVERALL ACCURACY + or – a Rank	0.72 - 0.94

Several models were applied whereby multiple look-back periods were used as inputs in addition to the position relative to the tanker. Next, classification models were built across each behavior. The classifiers were then applied to test data to review the accuracy at detecting qualifying events.

Figure 7. Event Triggering Model

Unfortunately, the models faced a setback in detecting qualifying events, signaling a need for richer data inputs to address this aspect effectively. While we were able to apply a satisfactory rating for a given behavior, we were unable to detect when a given behavior event had occurred. To better accommodate near real time event-based predictions, future data could be improved through features such as user identified critical event windows (i.e., 5 second lead up to an event) or predefined behavior flags (i.e., user moved far left). This event-based predictive framework holds promise beyond the realm of air refueling, offering versatile applications across various user-ranking dependent time

domain use cases. The scalability of this conceptual framework to diverse contexts, including step-based procedural training scenarios, opens doors to broader utility and effectiveness.

Continuing our exploration, we pursued a model geared towards determining performance ratings solely from simulator data. We aggregated all the training data over the entirety of the session to look for any relationships to the behavioral score. We considered this approach to measure the aircraft handling of the user for that session. The model inputs included min, max, mean and variance measures across the virtual flight inputs. We then built a model that was able to fit the existing data. K-fold cross-validation was used to determine how well the model generalizes on new data. The model did not generalize well enough to determine a satisfactory rating at the 1 through 5 level, however, it did show that this approach has strong evidence of providing valuable information in qualifying good and bad sessions. Future opportunities to enhance the model are discussed in more detail below. We investigated whether the model could generalize a rating lower or greater than 2.5 rather than provide a specific rating. The 2.5 threshold was chosen because it provided clean delineation of higher and lower performance.

Within the limited initial data set, the model was able to generalize with an accuracy of 80% (0.8 recall, 0.8 precision, 0.8 accuracy) having 1 false negative and 1 false positive detected. The inputs that were used included aggregations of the position deltas across all the x, y and z planes between the tanker and the plane during all the connected states observed in a session. That is to say, the min, max, mean and variance positions between the tanker and the aircraft over an entire session could be used to rate the aircraft handling of the session. This model was unable to successfully predict when an event occurred despite being able to accurately (70-90%) assign a rating for that event. Ultimately, it was found that the model could not generalize well enough to determine a satisfactory individual rating on performance but proved valuable at the overall level (i.e., generalizable above and below the 2.5 threshold to determine ‘good’ or ‘needs improvement’ for overall performance). In future, given additional data sets, it is possible to expand this assessment to multiple proficiency levels (i.e., good, better, best), or to modify the threshold for proficiency (i.e., from 2.5 to 3.5 as the threshold for ‘mastery’).

This model was implemented as ‘Automated Performance Insights’ within the training platform, which classifies overall performance as ‘good’ or ‘needs improvement’ at the session level. Outcomes demonstrate the potential to provide automated assessment in the absence of behavior ratings and has the potential to be scaled to new user cases and training environments in the future with further development.

FUTURE OPPORTUNITIES

Model Enhancements

Outcomes of the current research resulted in an automated performance assessment approach that utilizes representative flight data to classify users as either 'good' or 'needing improvement' with 80% accuracy. This method was selected for its high accuracy and potential for expansion with additional datasets in the future. Beyond the current work, we have investigated potential enhancements such as incorporating additional features, refining the threshold for proficiency levels, or exploring more advanced modeling techniques to enhance predictive accuracy. One such example is to provide adaptive feedback based on high-level performance categorization as the training system matures. This feedback could include contextualized recommendations such as tips or prompts to address specific performance gaps. Although real-time event triggers or recommendations require two-way communication between the virtual trainer and the prototype system—a capability not yet supported within the primary use case—there is significant value in maintaining a repository of categorized recommendations to offer technique tips based on identified skill gaps after each session. For instance, advising a student to adjust their arm and hand positioning on the throttle for more precise control when they receive low ratings on throttle-related behaviors. Future advancements in trainer technology could facilitate a broader range of real-time adaptive feedback. Additionally, changes in the types of data outputs produced will enable support for event-based performance models currently being explored.

Organizational Analytics

Currently, the resulting capabilities provide contextually-rich results through automated, AI–assisted performance review paired with behavior-based ratings and feedback at the individual session level. There is an opportunity to further leverage the existing architecture to support large scale performance tracking and analytics to identify skill gaps organization wide through customizable data visualizations. Current capabilities include the ability to view

analytics through organizational metrics, cross comparisons, and user profiles. The cross comparisons feature supports filtering and visualization of performance data across individual learners, groups, and observers. Post hoc groups can be created and comparisons between individuals and groups at specific instances and over time are possible. This feature allows administrators to track how the performance of individuals and groups compare over specified durations of time (e.g., the duration of a course) across a wide range of data visualization. This is currently possible at the behavior level, but expanding to include APMs within the analytics capabilities would allow users and instructors to track changes in performance across individuals and groups overtime. This capability is especially valuable as there is a need to quickly validate the effectiveness of the new and emerging simulation training technology being developed and acquired DoD wide**.**

Gamification of Results

There is significant potential to apply gamification techniques—using game design elements in non-game contexts—within military simulation training environments. According to Kim et al. (2018), the application of gamification in learning and education can result in several positive outcomes, including increased engagement and motivation, improved retention, and more contextualized feedback and performance review. Our domain analyses with SMEs indicate that student pilots thrive in both collaborative and competitive learning environments. Incorporating specific game features such as leaderboards, personal best scores, and individual and group achievement awards can enhance self-guided learning and foster interaction, ultimately leading to accelerated learning and increased retention.

CONCLUSION

In conclusion, the integration of advanced technologies into military training, particularly through the use of virtual and simulation-based systems, offers immense potential for enhancing training outcomes. This research focused on developing an adaptive training engine that leverages validated training interventions, such as behavior-based performance assessment, video self-modeling, self-reflection, and personalized feedback. By incorporating machine learning techniques, this engine can analyze data from a virtual trainer and integrate instructor ratings to deliver cohesive after-action reviews enriched with AI-driven performance insights.

The use case of B-52 air refueling highlighted the critical need for a streamlined after-action review process in simulation-based pilot training. The development of a contextualized rating rubric, based on Behaviorally Anchored Rating Scales (BARS), provided a systematic framework for evaluating performance and enhancing rater objectivity. This approach enabled the identification of key performance areas and behaviors critical to the air refueling task, ensuring a robust and equitable assessment mechanism.

Through adaptive performance modeling and the implementation of automated performance insights, the research demonstrated the feasibility of reducing the burden on human instructors while enhancing the granularity and accuracy of performance evaluations. Although challenges remain in detecting specific behavior events and providing real-time adaptive feedback, the findings underscore the need for continued refinement and the potential for future improvements with additional data sets and advancements in trainer technology.

Overall, this research underscores the significant advantages of integrating virtual training systems with advanced assessment tools to improve the readiness and adaptability of warfighters. By providing detailed performance feedback and leveraging data analytics, military training can become more efficient, effective, and responsive to the evolving needs of modern combat environments. The development of such adaptive training engines holds promise for large-scale performance tracking and organizational analytics, further enhancing the strategic capabilities of military forces.

REFERENCES

Dowrick, P. W. (1999). A review of self-modeling and related interventions. Applied and Preventive Psychology, 8(1), 23-39.

Dreyfus, H. L. and Dreyfus, S E (1986) Mind over Machine: the power of human intuition and expertise in the age of the computer, Oxford, Basil Blackwell

Fadde, P., & Sullivan, P. (2013). Using interactive video to develop preservice teachers' classroom awareness. Contemporary Issues in Technology and Teacher Education, 13, 156-174.

Fiorella, L., & Vogel-Walcutt, J. J. (2011, September). Metacognitive Prompting as a Generalizable Instructional Tool in Simulation-Based Training. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 55, No. 1, pp. 565-569). SAGE Publications.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. Journal of applied psychology, 83(2), 218.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112.

Kim, S., Song, K., Lockee, B., & Burton, J. (2018). Gamification in Learning and Education: Enjoy Learning Like Gaming: Springer.

Lele, A. (2013). Virtual reality and its military utility. *Journal of Ambient Intelligence and Humanized Computing*, *4*, 17-26.

Malone, N., Vogel-Walcutt, J., Ross, K., & Phillips, J. (2014). "Literature review of competencies and standards for instructor development." Program Technical Report for ONR Contract Number N00014-14-C-0106.

McDale, S., & Ma, J. (2008). Effects of fatigue on flight training: A survey of US part 141 flight schools. *International Journal of Applied Aviation Studies*, *8*(2), 311-336.

Stanney, K. M., Archer, J., Skinner, A., Horner, C., Hughes, C., Brawand, N. P., ... & Perez, R. S. (2022). Performance gains from adaptive eXtended Reality training fueled by artificial intelligence. *The Journal of Defense Modeling and Simulation*, *19*(2), 195-218.

Wu, L., & Looi, C. K. (2012). Agent Prompts: Scaffolding for Productive Reflection in an Intelligent Learning Environment. Educational Technology & Society, 15(1), 339-353.

Yu, J., Lo, C., Claudia, M., Jagmeet, B., Olszynski, P., & Malcolm, L. (2020). Video modeling and video feedback to reduce time to perform intravenous cannulation in medical students: a randomized-controlled mixed-methods study. *Canadian Journal of Anesthesia*, *67*(6), 715-725.