# From Innovation to Integration: Data-Driven Evaluation of Modernized Training

**Alexxa Bessey, Summer Rebensky, Mark Shroeder-Strong**
**Aptima, Inc.**
**Woburn, MA**
**abessey@aptima.com, srebensky@aptima.com, mschroeder@aptima.com**

**Brian Schreiber**
**Crystal Lake, IL**
**schreiber296@aol.com**

**Steve Macut**
**BGI, LLC**
**Akron, OH**
**Steven.Macut@bgi-llc.com**

**Winston "Wink" Bennett**
**Bennett Research Consulting LLC,**
**Lawrenceburg, KY**
**bennettresearchconsultingllc@gmail.com**

## ABSTRACT

In the contemporary landscape of military training, the reliance on simulator-based training and the introduction of novel training technologies has surged, promising both enhanced efficiency in training warfighters and a competitive edge when facing adversaries. However, amid this influx of innovation, a critical void remains: the absence of robust methodologies to assess the integration and impact of these training technologies on relevant outcomes (e.g., readiness, proficiency, performance). Stated another way, a systematic approach is needed to demonstrate that these training technologies provide adequate or better training than the legacy simulators and live training they are augmenting and replacing. Further, there lacks guidance on where, when, and how to implement new training technologies. This lack of guidance can impact the acquisition of training technology, development of training curricula and instruction, and optimization of training requirements. Even with current ad-hoc methods, military personnel lack a data-driven approach to establishing new requirements and providing actionable feedback to vendors on training technology deficiencies. As a result, the purpose of the following paper is two-fold. First, the following paper will discuss the importance of a data-driven understanding when integrating and evaluating the impact of new training technologies. Next, the following paper will provide data and recommendations from three case studies that have systematically evaluated and compared different training technologies within the United States Air Force (USAF) and efforts towards modernizing the Joint Simulation Environment (JSE). Through this analysis, we advocate for a paradigm shift towards a comprehensive and systematic evaluation approach that prioritizes not only technological advancement but also the effective integration and utilization of these technologies within military training to guide future requirements. By addressing this critical gap, we aim to catalyze actions that prioritize data-driven decision-making as well as discuss best practices in an era defined by rapid technological evolution.

## ABOUT THE AUTHORS

**Dr. Alexxa Bessey** is Scientist in the Training, Learning, and Readiness Division at Aptima, Inc. She has over 7 years of experience conducting research within military environments, including her time spent as an operational research psychologist in the field. Dr. Bessey offers an expertise in simulator assessment, unobtrusive measurements, training, and teams. In addition, she is well-versed in data collection and sampling methodologies in military settings, to include both subjective and objective approaches. At Aptima, Dr. Bessey is involved in several projects including the assessment of proficiency-based training, the evaluation of simulator fidelity, and the examination of unobtrusive measures in teams. As part of her work at Aptima, in addition to her doctoral work, Dr. Bessey's research examines validating unobtrusive data measurements. Dr. Bessey has a master's degree in both Clinical Psychological Science and Industrial-Organizational Psychology and a PhD in Industrial-Organizational Psychology from Clemson University. US Citizen.

**Dr. Summer Rebensky** is a Scientist at Aptima and has expertise focusing on human performance, cognition, and training in emerging systems. In her role at Aptima, Inc. she serves as the capability lead for Air Force Training, Learning, and Readiness technologies. She specializes in leading efforts to develop, test, and implement AR, VR, XR, and modern training solutions to improve and measure human performance. Dr. Rebensky has previous experience as a research fellow as a part of the Air Force Research Laboratory's Gaming Research Integration for Learning Laboratory (GRILL) conducting research on interface designs for novel aviation use cases in the civilian and DoD world as well as and human-agent teaming designs. Her other experience includes leading Florida Tech's ATLAS research lab with grant and STTR research with the Air Force, Navy, and FAA; human factors assessments of a trainer aircraft for the F-35 with Northrop Grumman; and developing DoD courseware with Raytheon. Her research experience involves leveraging VR, AR, and modern technology to optimize human performance in training and operations, individualizing learning through gamification and adaptive technologies, human-agent teaming, and trust in AI. Dr. Rebensky received her BA in psychology, MS in aviation human factors, and PhD in aviation sciences with a focus in human factors from Florida Tech. US Citizen.

**Brian Schreiber** is a Principal Scientist. He holds a master's in science from the University of Illinois at Champaign-Urbana and has been performing research and development within the military domain since 1993. His research has focused on developing and evaluating military training techniques and technologies. Some of those projects include developing the Performance Evaluation Tracking System (PETS) and the first versions of SimMD, evaluating skill acquisition/decay, evaluating the return on investment of training environments and individual training aids (e.g., motion seats, eye trackers), developing/expanding standards, and serving as a consultant. He has authored or co-authored over 65 papers, journal articles, tech reports, and book chapters. US Citizen.

**Dr. Mark Schroeder-Strong** is a Research Scientist at Aptima, Inc. and has more than 13 years of experience in the field of applied training effectiveness research. He is also an Associate Professor of Educational Foundations at the University of Wisconsin–Whitewater, where he teaches courses in measurement, teacher education, and development. Over the past decade, he has conducted research examining skill decay, the impact of fidelity enhancements on training effectiveness, training capability assessment techniques, and the organization and application of automated data collection for objective performance measures. Dr. Schroeder-Strong has played a significant role in the initial design and extension of the capabilities of Sim MD, an SBIR Phase II-funded technology that facilitates networked evaluations of training systems to document capabilities, identify deficiencies, and provide a path toward improvements. He was also Co-Principal Investigator on an SBIR Phase II-funded program, Predicting, Analyzing, and Tracking Training Readiness and Needs (PATTRN), which improves training programs by tracking trainee proficiencies, predicting future training needs, and providing instructors with recommendations and organizational tools to deliver just-in-time training. Dr. Schroeder-Strong's academic interests lie in exploring how causal relations impact perceptions and policy in education. He holds a PhD in educational psychology from the University of Wisconsin-Milwaukee. US Citizen.

**Steven Macut** is a Senior Operational Analyst with BGI. He has 3 years of experience supporting Air Force Research Lab's Proficiently Based Training initiative. He has a bachelor's degree in Aerospace Engineering from Penn State University. He is retired Air Force with 15 years of experience flying the F-15E. Additionally, he has 12 years of experience as a Simulator and Academic Instructor at the F-15E Formal Training Unit. US Citizen.

**Dr. Winston "Wink" Bennett** received his Ph.D. in Industrial Organizational Psychology from Texas A&M University in 1995. He has been involved in NATO-related research activities for over 20 years. He recently retired from AFRL and is now supporting the DAF Chief Modeling and Simulation Office with SAIC given his previous work in modeling and simulation standards development, test and evaluation of live, virtual and constructive interoperable and distributed training technologies, and the application of novel technologies to address USAF local and deployed readiness requirements. He has also been involved in I/ITSEC committee and program work for a number of years as well. He maintains an active presence in the international research and practice community through his work on various professional committees and his contributions in professional journals and forums including I/ITSEC. His involvement with the larger psychological communities of interest ensures that communication amongst international military, industry and academic researchers remains consistent and of the highest quality. US Citizen.

# From Innovation to Integration: Data-Driven Evaluation of Modernized Training

**Alexxa Bessey, Summer Rebensky, Mark Shroeder-Strong**
**Aptima, Inc.**
**Woburn, MA**
**abessey@aptima.com, srebensky@aptima.com, mschroeder@aptima.com**

**Brian Schreiber**
**Crystal Lake, IL**
**schreiber296@aol.com**

**Steve Macut**
**BGI, LLC**
**Akron, OH**
**Steven.Macut@bgi-llc.com**

**Winston "Wink" Bennett**
**Bennett Research Consulting LLC,**
**Lawrenceburg, KY**
**bennettresearchconsultingllc@gmail.com**

## INTRODUCTION

The contemporary landscape of military training presents as a dynamic and rapidly evolving blend of traditional methodologies and cutting-edge technologies. Historically rooted in live exercises and classroom instruction, military training has undergone a transformation with the introduction of simulator-based training and the integration of innovative training technologies. Such advancements have revolutionized the training paradigm, offering immersive and realistic experiences that simulate operational environments with unprecedented fidelity and adaptive learning. In addition to the changing nature of training technology, the shifting geopolitical landscape and evolving threats necessitate agility and adaptability in training approaches. Taken together, there is a compelling case for the military to embrace innovation and leverage emerging technologies to ensure the readiness and effectiveness of warfighters in increasingly complex and dynamic environments.

Despite the rapid advancement and widespread adoption of simulator-based training and innovative training technologies, there lacks implemented evaluation methodologies that can comprehensively assess the impact of newly introduced training technology on crucial training outcomes. Stated another way, it may be insufficient to simply introduce new technologies into existing training paradigms without better understanding how they can meet existing training requirements as well as their influence on training outcomes such as readiness, proficiency, and performance. In addition, without methodologies and test and evaluation (T&E) data, it becomes difficult to ascertain whether these technologies deliver superior or even comparable training to traditional and existing training approaches. On the contrary, it could be that while more advanced, the emerging and innovative training technologies may be deficient or even counterproductive to meeting training goals. This notion, that innovative training technology may not always deliver superior or even productive training, is highly relevant not only to warfighter performance but also acquisition processes and curriculum development. More specifically, having standardized T&E data from new training technology are needed to guide strategic decision-making in technology acquisition prior to the deployment of technology to ensure efficiency and effectiveness when training technology is procured. Further, once the training technology is deployed, T&E data should continue to be collected to provide actionable feedback to technology vendors when training deficiencies emerge. As both the training and technological impacts become realized, T&E data should also be utilized to inform the development of updated curriculum. In summary, T&E data are crucial at several points of the training pipeline in order to ensure training technology optimization and warfighter readiness.

With a dual-fold purpose, the following paper aims to highlight a critical gap within military training and advocate for a data-driven approach to evaluating the integration and impact of emerging training technologies. By emphasizing the importance of systematic evaluation methodologies, the paper advocates for a data-based foundation for informed decision-making and strategic planning in the adoption and implementation of these technologies. Secondly, the paper seeks to provide actionable insights, recommendations, best practices, and lessons learned from three comprehensive case studies. The described case studies each cover a unique systematic evaluation and comparison of different training technologies. Ultimately, the overarching objective of the paper is to shift the field towards a more comprehensive and systematic evaluation approach which can better enable the use of training technologies and improve training outcomes.

### Innovation and Evaluation

Technological innovation has significantly reshaped military training, enhancing effectiveness and adaptability in modern warfare scenarios. Advanced simulators, augmented reality (AR), virtual reality (VR), and artificial intelligence (AI) have revolutionized training paradigms, offering realistic and dynamic environments for warfighters to sharpen their skills. These innovations provide immersive experiences that replicate combat situations, allowing trainees to develop decision-making capabilities and muscle memory in a safe yet realistic setting. Stated another way, training technologies are often able to offer training experiences that would otherwise be difficult to train to in live training environments, such as emergency procedures (Myers et al., 2018). According to a report by the Congressional Research Service, integrating cutting-edge technology into military training programs not only enhances effectiveness but also reduces costs and risks associated with live exercises (Sayler, 2022; Sayler, 2024). Furthermore, technologies like AI-driven training systems can analyze performance data to tailor individualized training regimens, optimizing the learning process. In a military environment, where preparedness and quick decision-making are critical to success, such advancements play a crucial role in ensuring readiness and operational success.

While innovation is important, the use of a systematic evaluation methodology is also essential in that it provides both an assessment framework that ensures that all relevant factors are considered and evaluated as well as T&E data. Much like other types of training evaluation, the advantages of data-driven decision-making in evaluating training technology impacts can be multifaceted and profound. First, leveraging data allows for an objective and evidence-based assessment of the effectiveness and efficacy of training technologies. By analyzing quantitative and qualitative metrics, decision-makers can gain valuable insights into the real-world impact of these technologies on training outcomes. For example, the data can be used to extrapolate strengths, weaknesses, and areas for improvement of the training technology. Additionally, data-driven approaches enable the identification of trends, patterns, and correlations that may not be apparent through ad hoc or informal feedback. Furthermore, data-driven decision-making fosters accountability and transparency, as conclusions and recommendations are grounded in empirical evidence rather than subjective opinions or biases. Ultimately, by harnessing the power of data, military organizations can make more informed and strategic decisions regarding the integration, optimization, and utilization of training technologies, thereby maximizing their impact on important outcomes.

While the advantages of of data-driven evaluation approaches are important to highlight, equally important to discuss are the implications of the lack of robust methodologies for training technology assessment. To begin, without standardized and comprehensive evaluation frameworks, there is a risk of inconsistent, unreliable, and overinflated assessments of training technology effectiveness. If done at all, often these assessments are done with small sample sizes and/or are gathered informally. This lack of data or inconsistent data can lead to disparate conclusions regarding the utility and impact of these technologies. This can hinder informed decision-making and planning efforts during the training technology acquisition and deployment process. Moreover, the absence of robust methodologies impedes efforts to optimize and improve training due to the lack of data needed to draw conclusions with confidence. In many cases, the small or informal sample sizes are not sufficient for decision making. Lastly, the lack of standardized assessment methodologies may result in missed opportunities to identify emerging trends or technological advancements that could further enhance training effectiveness. Stated another way, T&E data can help to identify innovative training technologies that provide enhanced training, allowing for the military to make more significant investments in technology with a greater return on investment (ROI).

## CASE STUDIES

### Overview

The following three case studies provide examples of the evaluation and comparison of different training technologies. These case studies offer valuable insights into the practical, systematic evaluation of both existing and emerging training technologies within real-world military training environments. Through these case studies, the paper aims to illuminate best practices, identify key success factors, and highlight potential areas for improvement in systematically evaluating training technologies within military training paradigms.

Each of the three case study evaluations used the same systematic process to collect feedback. This evaluation process offers a highly adaptable framework applicable across many different domains and training devices, as demonstrated by the diversity of evaluated training technology and environments across the three case studies. Although each evaluation is slightly different, the evaluation content is often derived from established training documentation such as Training Task Lists (TTLs) and Ready Aircrew Program (RAP) Tasking Memos (RTMs). The content then undergoes customization in collaboration with key stakeholders and subject matter experts (SMEs) to ensure relevance and alignment with training requirements. Subsequently, a comprehensive set of primary and secondary deficiencies is developed to categorize areas of technological inadequacy. The deficiency sets are structured such that each primary deficiency (e.g., Visuals) has a sub-set of more specific, corresponding secondary deficiencies (e.g., Resolution, Field of View, Display). The use of primary and secondary deficiencies allows for trend analysis and identification of key areas for improvement. During the evaluation, operators utilize a Likert scale to rate each evaluation item, with the option to select primary and secondary deficiencies for items rated as deficient in supporting training. When the evaluations include comparisons, evaluators are asked to rate each training technology for each evaluation item. They are also asked to provide detailed open-ended comments for items rated as deficient. The use of Likert scale data, deficiency data, and comment data creates a multimodal dataset that can be used to extract different insights. The raw data are then exported as a JavaScript Object Notation (JSON) file and compiled into an Excel report template. The report provides a comprehensive overview of the training environment's quality, including inter-rater reliability, strengths, areas of improvement, potential training gaps, participant demographics, and summaries of comment data.

## Case Study #1: USAF Command and Control Platform

Within simulated training environments, certain elements of the virtual and simulated environment can provide varying levels of training. As new technology and training modalities are introduced, evaluations must occur to examine their impact on training. This is especially important as training environments may be different across units. As a result, the first use case includes an evaluation of a USAF Command and Control (C2) platform. The C2 platform was evaluated to better understand the extent to which different training modalities provided adequate training of documented requirements. More specifically, the evaluation examined the source of training, comparing locally executed training with training administered by a distributed training center through the distributed mission operations network (DMON). Consequently, the evaluation was structured to assess existing training requirements within local, standalone simulator training and distributed simulator training. Stated another way, evaluators provided feedback on their ability to train to an established training requirement (e.g., defensive counter air) when training locally vs. when they are training through a distributed training center. Each training modality was examined individually as well as comparatively. The C2 evaluation occurred across four sites and included both CONUS and OCONUS units as well as active-duty and national guard units. Across the four sites, a total of 152 operators were evaluated. Given the variety of platform positions with unique tasking, two different evaluations were administered. Evaluation group 1 (N=65) included positions part of Surveillance/Links operators. Evaluation group 2 (N=87) included positions part of Weapons Section/Position operators.

For both evaluation groups, evaluators rated their ability to meet their training requirements in the distributed training environment higher than in the local training environment. Similarly, the average rating was higher in the distributed training environment than in the local training environment. However, only evaluation group 2 had statistically significant differences between the average distributed training item and the average local training item. Stated another way, evaluation group 2 rated their ability to meet their training requirements significantly better in the distributed training environment than in the local training environment (See Figure 1).

The training environments were further examined to better understand training differences at the requirement level to identify the optimal training environment for each training requirements. For evaluation group 1, 5% of training requirements were demonstrated to be effective in the distributed training environment but were found to be deficient in the local environment. Similarly, for evaluation group 2, 15% of training requirements were demonstrated to be effective in the distributed training environment but were found to be deficient in the local environment. Stated another way, the evaluation of the two environments revealed that a portion of training requirements should be trained in the distributed environment opposed to the local environment as the local environment is deficient in providing adequate training. Conversely, one training requirement for evaluation group 2 was demonstrated to be effective in the local training environment but was found to be deficient in the distributed environment. In this case, it was recommended that this training requirement should only be trained locally due to deficiencies within the distributed environment. Lastly, 7% of the training requirements for evaluation group 2 were found to be deficient in both training environments. Stated another way, both the local and distributed training environments failed to provide sufficient training. As a result, improvements to the training environments should be made to ensure operators are able to meet the training requirement in at least, if not both, environments.

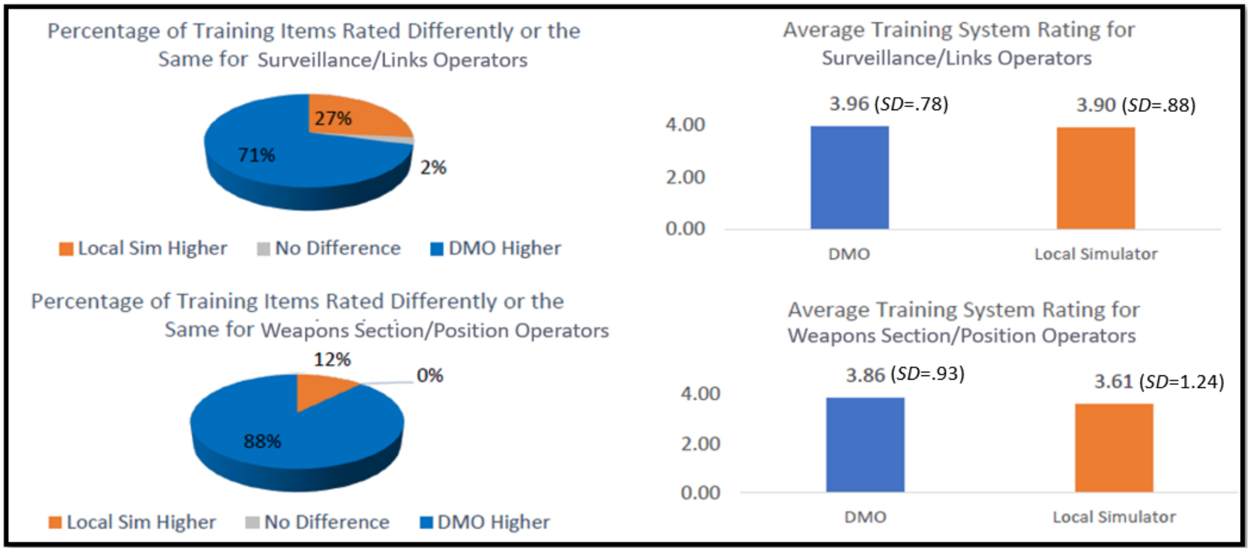

**Figure 1. Case Study #1 Results**

In addition, deficiencies were examined for both training environments. Across both groups, the distributed training environment had 16 training requirements that were rated as deficient by at least one site. The local training environment had 35 training requirements that were rated as deficient by at least one site. Deficiencies were most often related to the scenarios utilized during training (e.g., scenario realism, scenario variety), IOS (e.g., IOS implementation), connectivity, and mission preparation. Comment data were analyzed for trends. For the distributed training environment, operators noted a lack of clarity on training center capabilities, limitations with training center scenario development, and logistic issues related to scheduling and connectivity. For local training, operators noted the lack of locally based resources to execute training which taxed other aspects of training. For example, the necessitation of instructors to drive training, gaps in understanding how to use local simulation training tools, and the lack of debrief capabilities. From this evaluation, insights can be drawn regarding the extent to which both distributed and local training can train to existing requirements. For evaluation group 2 specifically, this evaluation highlighted several concerns regarding their training, explicitly as it relates to training done locally. Other key insights from this evaluation included the identification of training requirement strengths within each environment, training requirements that were adequately trained regardless of environment, and training requirements that were not being met in either environment. Such insights are critical for leveraging training environments to optimize training. It is also important for understanding where resources should be invested. In addition, comment and deficiency data were analyzed in order to identify key trends that were impacting the quality of training. From that trend analysis, highly specific recommendations were able to be provided.

A key lesson learned from this evaluation was understanding the differences in training environments across sites. Although all four sites for the C2 platform had the same established training requirements, each site had nuanced differences that impacted the training environment. For example, the two active-duty sites had additional personnel that could assist with aspects of both local training as well as connecting to the DMON during distributed training. In another example, one site had a particularly skilled individual who had increased understanding of navigating the local training tool, to include the development of more robust scenarios. As a result, the ability to train to the established training was impacted differently across the four sites. However, despite cross-site differences, several training requirements were demonstrated to be sufficient and deficient across all four sites. Such observations further highlight the need for both aggregated and site-specific empirical data that can speak to and identify strengths and weaknesses that are specific to a singular site and those that are more widespread across the force. This is especially important for assessing readiness. In summary, the described case study provides several examples of how systematic data can be leveraged to inform the improvement of training outcomes. Further, it underscores the notion that not all training environments or training sites are equal and understanding such differences is key to optimizing training and ensuring readiness.

**Case Study #2: USAF Platform Iterative Design Assessment**

Data-driven evaluation approaches can be used for test and training centers to understand novel system design current gaps and improvements needed. Many technologies that are tested in operational mission sets but are still not up to snuff for prime time. Some technologies are tested in iterations with consistent revisions. These are referred to as evolutionary acquisition processes—a preferred method of the Department of Defense (DoD) for rapid acquisition where the capability can be put in front of the warfighter sooner, with continuous development occurring over time as a collaboration between user, tester, and developer (USAF TE, 2019). Historically, test reports cover the insights of the team testing the technology. Over multiple pages, test procedures and test personnel insights are described. The results of tests related to pain points, performance (e.g., network latency) and other metrics may be presented. However, these assessments may lack quantitative data from a range of end users with priority focused debrief that can quickly give decision makers insights. Leadership needs actionable insights into the necessary feedback to give to vendors, or requirements to right into the next period of performance's work statement. Test and training centers attempt to deliver these actionable insights: "As painstakingly detailed and time-consuming as they often are, test reports often mask the reality that the data collected is insufficient and the conclusions drawn are simply a semi-informed best guess. To rapidly deliver warfighter capability, the DoD must accept increased risk and accept fielding recommendations based on the reality of limited available data" (Bradley, 2022). Many do not have the resources to conduct full scaled studies. However, in the case of evaluating iterative technology, a majority of the problems can be uncovered in as little as 5-16 individuals (Neilsen et al., 1993). Leveraging an evolutionary acquisition model and testing early and often can afford leaps and bounds in providing the warfighters improved technology sooner. This evaluation method was implemented with a test and training center to evaluate a new virtual reality solution. As an initial trial of data collection in this domain, a total of seven pilots evaluated a simulation platform in multiple simulation scenarios in a JSE-like environment.

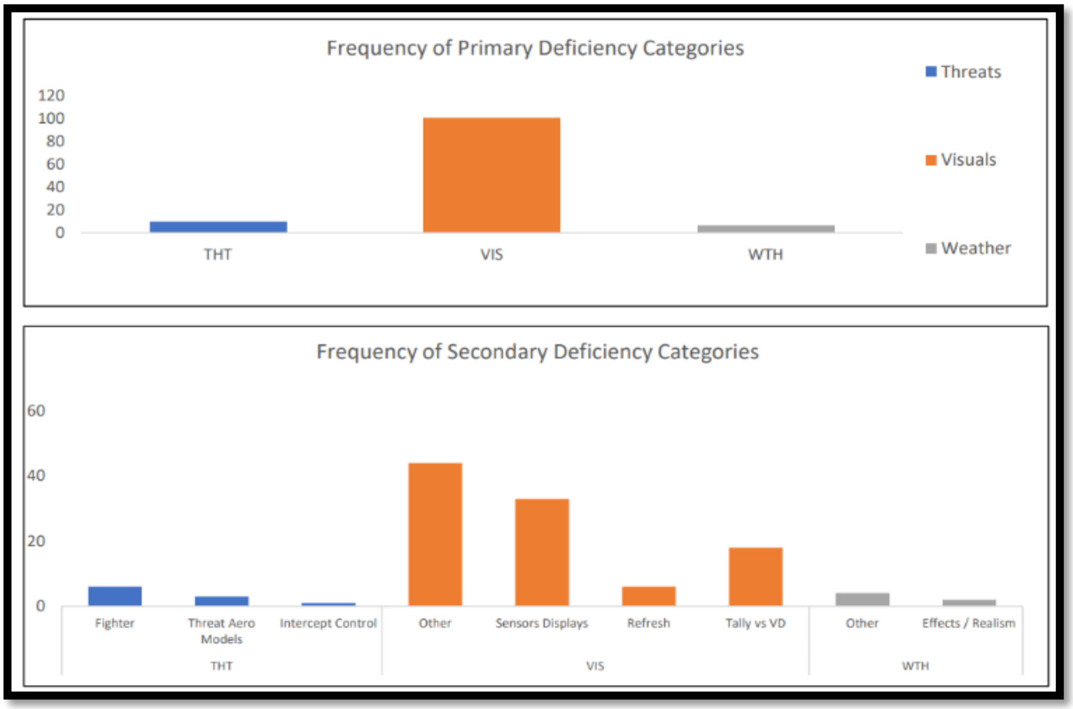

**Figure 2. Case Study #3 Results**

Pilots tested an extended reality (XR) simulation platform in representative operational scenarios. Pilots rated the ability for the system to enable specific mission skills such as judging the range of other aircraft and the performance of passthrough cameras on a scale from Deficient, Fair, Good, and Excellent. If a mission skill was rated as Fair or lower, the pilot would select a deficiency from categories of threats, visuals, or weather (see Figure 2). As seen in the figure, visuals were the largest category with noted issues. Comments uncovered issues with focus issues, field of view limitations, refresh rate, and visual identification. Multiple comment responses mention difficulty identifying threats over 1 mile out. The analysis allows the leadership to identify whether or not the system is ready to support the mission skills of interest. In this use case, results that demonstrate a good or excellent rating for instrument related procedures but ineffective for combat related procedures would align the solution for basic flight training. However, the leadership could choose to allocate the solution for that training use case or return to the vendor for the next spiral of the evolutionary acquisition specifically targeting the improvement of the visuals. If the vendor would be unable to make improvements in that area, then leadership would have data driven insights for standing up new competitive solicitations for solutions with clear and actionable requirements (e.g., the resolution of the display must be able to support visual identification of threats at one mile away). Towards later spirals of evolutionary acquisition, the leadership can decide to finalize the process once all necessary mission elements receive favorable ratings.

**Case Study #3: USAF Command and Control Capability Assessment**

The third case study is an upcoming application. As the DoD moves towards joint all domain command and control (JADC2), it is paramount that technologies help enable faster information processing, synthesis of information, decision support, and actions in the battlespace (DoD, 2022). There are a myriad of solutions that could assist joint services in assess the battlespace, determining battle strategies, and assisting the operators are enormous. Improvements to command and control must be targeted in scope to support the JADC2 objectives as soon as possible. However, deciding which technologies, tactics, techniques, or procedures (TTPs) that can benefit first and foremost from investment is a tough answer to find and difficult to assess the best solution. The rise of multiple technologies in the C2 space has demanded the efforts of transformational models (TM) which use model-based systems engineering to distill C2 functions within the domain. One of the biggest motivators being "DAF leadership requires evidence-based acquisition strategies" (ABMS CFT, 2023). This methodology, a form of mission engineering (ME), is specifically designed to address the complexities of modern and future military operations under the Combined Joint All-Domain Command and Control (CJADC2) framework. Key goals of the Transformational Model include quantifying specific functional and high-level non-functional requirements for Joint and Combined Operations with an emphasis on C2 decision functions, categorized into battle management (TM-BM), planning (TM-P), and command (TM-C) decisions. It aims to improve DoD decision making on changes to TTPs by decomposing C2 decision-making into granular enough levels to understand. This will enable measurable impacts of Human-Machine Team (HMT) decision function performance through technological and other DOTMLPF-P (Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, Facilities, and Policy) solutions.

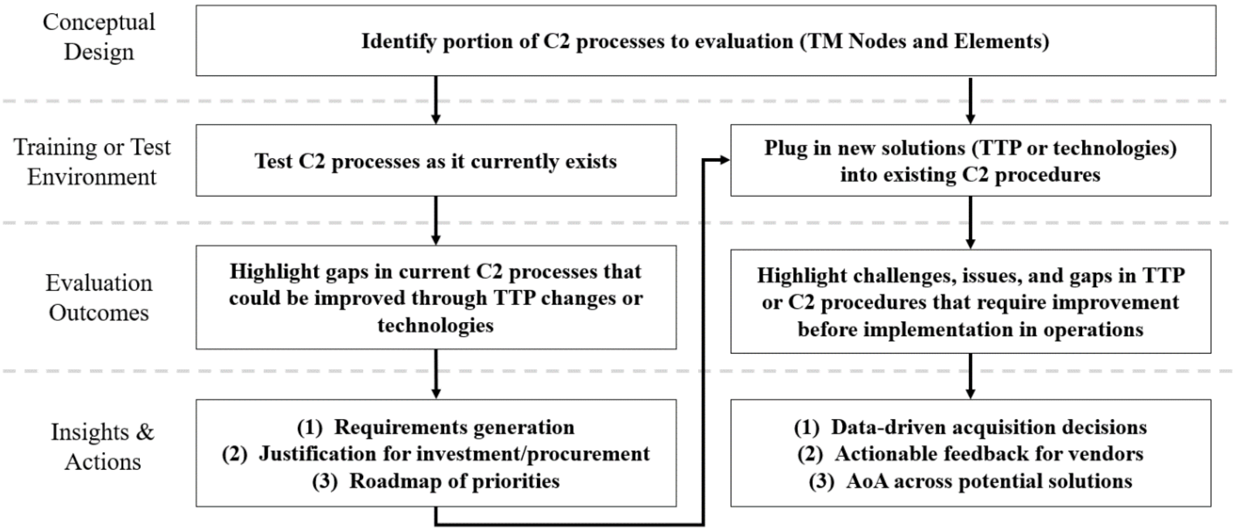
Conceptual Design
Identify portion of C2 processes to evaluation (TM Nodes and Elements)
Training or Test Environment
Test C2 processes as it currently exists
Plug in new solutions (TTP or technologies) into existing C2 procedures
Evaluation Outcomes
Highlight gaps in current C2 processes that could be improved through TTP changes or technologies
Highlight challenges, issues, and gaps in TTP or C2 procedures that require improvement before implementation in operations
Insights & Actions
(1) Requirements generation
(2) Justification for investment/procurement
(3) Roadmap of priorities
(1) Data-driven acquisition decisions
(2) Actionable feedback for vendors
(3) AoA across potential solutions

**Figure 3. Analysis application to C2**

The same approach utilized in case study 1 & 2 can be applied to achieve the goals for case study 3. Each aspect or node within the TM model can be rated and tested against the others. Each new technological or DOTMLPF-P element can be rated and flagged if current gaps or deficiencies are present. In collaboration with a DoD experimentation lab, the recommended changes and emerging innovative C2 strategies can be tested. New concepts and technologies are tested in real-world scenarios to ensure their efficacy and integration into broader military frameworks. The evaluation approach, targeting specifically the nodes and elements of the TM, can allow end users to highlight current gaps, identify areas in need of improvements, and understand the ripple effects that new TTPs or technologies have on the broader TM-BM, TM-P, or TM-C actions and decisions. Outcomes of such an analysis could help generate requirements for technology RFIs, generate funding requirements or a roadmap of priorities. Similar to case study 2, once new solutions are sought, they can be tested within DoD test and experimentation labs and evaluated again. This process ensures that the DoD can be more evidence based in acquisition processes and answer "Where can I best spend my dollar?" (see Figure 3). Considering the impact one decision can have on the broader mission, CJADC2 is likely to experience ripple effects from any small changes to the TTP. A similar approach to case study 1 could be implemented at multiple sites in a live, virtual, and constructive missions (e.g., Red Flag). One site could alter TTPs and the resulting effect on other sites could be observed. Additionally, multi-site outcomes could be analyzed and observed.

## CONCLUSION

Within military training, the fusion of traditional and innovative approaches is evident, with emerging training technologies reshaping and redefining existing training approaches. Despite this evolution, there is a critical evaluation gap that exists, risking both inefficiency and ineffectiveness within military training. More specifically, the introduction of new technologies into training paradigms without robust evaluation methodologies may be insufficient at ensuring training needs are met and key outcomes are not adversely impacted. Stated another way, without such comprehensive evaluation methodologies, discerning whether these technologies truly enhance training outcomes becomes challenging, potentially hindering real-world strategic decision-making in technology acquisition and training optimization efforts. As a result, the current paper serves as a call for data-driven decision-making, emphasizing the need for comprehensive, systematic methodologies for evaluating both novel training technology and the broader training paradigms they are used within.

In order to highlight the importance of systematic evaluation methodologies, three comprehensive case studies were discussed. The use of a systematic evaluation process, applicable across diverse training domains and technologies, provides a robust framework for evaluation as well as an opportunity to highlight best practices and key success factors. Looking ahead, future directions for advancing systematic evaluation approaches in military training include the use of standardized evaluation practices, consistent evaluation, and knowledge sharing. As argued in this paper, the military should move towards establishing standardized evaluation frameworks to ensure consistency and reliability in assessment outcomes. Such standardization will allow for cross-site comparisons of training technology and more widespread insights on impacts on training outcomes. Next, the evaluation process should be consistent to allow for longitudinal mapping of training outcomes as well as to inform different points of the training pipeline (e.g., acquisition, training, curriculum development). Lastly, a greater emphasis should be placed on knowledge sharing and collaboration among military organizations. An increase in knowledge sharing would allow the sharing of both insights from training technology evaluation (e.g., what technology worked best and in what environment) as well as the dissemination of best practices and lessons learned when evaluating raining technology. There are many military-wide use cases for the implementation of standardized evaluation practices, consistent evaluations, and knowledge sharing. However, such future directions are especially applicable as the military continues to stand up more complicated and interwoven training environments, such as the JSE.

In summary, a data-driven approach is crucial for maximizing the impact of emerging training technologies on military readiness. By embracing systematic evaluation methodologies and advancing future directions, military organizations can navigate the complexities of modern warfare with agility and efficacy, ensuring the maintenance of a highly capable and adaptive force.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the Air Force Research Laboratory and their partners for facilitating and supporting this effort as well as the warfighters who participated in the highlighted evaluations.

## REFERENCES

Air Battle Management Cross-Functional Team (2023). Transformational Model (TM) Methodology Core Concepts [Report].

Bradley, M. (2022). Accelerate test or lose. 53rd wing. https://www.53rdwing.af.mil/News/Article/3183005/accelerate-test-or-lose/

Department of Defense. (2022). Summary of the joint all domain command & control (JADC2) strategy [Report].

Myers, P., Starr, A., & Mullins, K. (2018). Flight Simulator Fidelity, Training Transfer, and the Role of Instructors in Optimizing Learning. *International Journal of Aviation, Aeronautics, and Aerospace, 5*(1). https://doi.org/10.15394/ijaaa.2018.1203

Nielsen, J., & Landauer, T.K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*.

Sayler, K. (2022). Military Applications of Extended Reality (pp. 1–3). https://crsreports.congress.gov/product/pdf/IF/IF12010/2

Sayler, K. (2024). *Emerging Military Technologies: Background and Issues for Congress* (pp. 1–39). https://sgp.fas.org/crs/natsec/R46458.pdf

United States Air Force Test & Evaluation (2019). Air Force Test and Evaluation Guide. [Report].