Simulators Provide Adequate Training – Says Who?

Alexxa Bessey, Brian Schreiber, Mark Schroeder	Steve Macut	Winston "Wink" Bennett
Aptima, Inc.	BGI, LLC	Air Force Research Laboratory
Woburn, MA	Akron, OH	Dayton, OH
abessey@aptima.com, bschreiber@aptima.com, mschroeder@aptima.com	Steven.Macut@bgi-llc.com	winston.bennett@us.af.mil

ABSTRACT

Simulated training environments have been identified as a vital training resource for the United States military. Compared to live training, such environments provide a cost-effective and safe alternative that can simulate a wide variety of training tasks, procedures, and exercises while minimizing the use of resources. With advancements in technology, simulated training environments now offer increasingly sophisticated training platforms that can be integrated and augmented with other technology, such as virtual reality and AI. However, as the military moves further into a digital training landscape and live training continues to be replaced by simulated training, it is key for leaders to understand the level of training fidelity simulated training provides. For example, inadequate training fidelity may be indicative of elements of simulated training that fail to provide training experiences compared to other training environments, such as live training or more robust, high-fidelity simulators. Understanding such deficiencies could inform areas of training that may need to be augmented and/or where to invest in improvements (e.g., upgrades to technology, scenarios, realism, etc.). Further, in addition to understanding inadequate training fidelity, it is also important to identify areas of adequate or exceptional training fidelity that can continue to be trained with a high degree of confidence. Simply put, understanding both the limitations and strengths of simulated training are critical for developing and sustaining well-trained warfighters. As a result, the following paper outlines a systematic approach to the evaluation of simulator fidelity that leverages subjective assessments of trainees as well as presents findings from a fidelity evaluation from a sample of United States Air Force (USAF) operators. In addition to findings from the fidelity evaluation, the following paper presents best practices and critical considerations when examining simulated training fidelity that are generalizable across the training community at large.

ABOUT THE AUTHORS

Dr. Alexxa Bessey is Scientist in the Training, Learning, and Readiness Division at Aptima, Inc. She has over 6 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.

Brian Schreiber is a Principal Scientist with Aptima. 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. He has authored or co-authored over 60 papers, journal articles, tech reports, and book chapters.

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.

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.

Dr. Winston "Wink" Bennett received his Ph.D. in Industrial Organizational Psychology from Texas A&M University in 1995. He is currently the Readiness Product Line Lead for the Airman Systems Directorate located at Wright Patterson AFB Ohio. He has been involved in NATO-related research activities for over 20 years. He has also been involved in I/ITSEC committee and program work for a number of years as well. He is spearheading the Combat Air Forces migration to proficiency-based training and is conducting research related to the integration of live and virtual training and performance environments to improve mission readiness and job proficiency. He leads research that has developed methods to monitor and routinely assess individual and team performance across live and virtual environments and evaluating game-based approaches for training, work design, and job restructuring. 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.

Simulators Provide Adequate Training – Says Who?

Alexxa Bessey, Brian Schreiber, Mark Schroeder
Aptima, Inc.
Woburn, MA
abessey@aptima.com, bschreiber@aptima.com, mschroeder@aptima.com

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

Winston "Wink" Bennett
Air Force Research Laboratory
Dayton, OH
winston.bennett@us.af.mil

INTRODUCTION

Increasingly, the Department of Defense (DoD) is becoming more reliant on simulated training environments (e.g., virtual reality, augmented reality, live-simulator blended environments) to provide training, as such environments offer several advantages when compared to traditional, live training experiences. For example, training devices such as simulators provide the ability to practice otherwise dangerous tasks, allow for multiple repetitions of a task, reduce costs, and minimize the expenditure of other resources (e.g., labor; Myers et al., 2018). Given the advantages of simulated training environments and the development of high-fidelity simulators, such simulators have been widely adopted by industries such as commercial airlines, medicine, and athletics as key sources of training (Ragan et al. 2015; Spencer, 2009). Within the United States Air Force (USAF), simulator-based training that mirrors platforms with varying degrees of fidelity are used to provide critical training while also addressing ongoing resource gaps within the USAF. For example, as a result of an aging fleet and increased energy costs, simulators have been used to provide low-cost training that minimizes the need for repetitive training events while also increasing training event effectiveness (Spencer, 2009). Although there are several advantages to simulators and simulator-based training that have been demonstrated empirically, there remains a gap in the effectiveness of applied approaches that examine fidelity as it relates to meeting the training requirements that are necessary for a well-trained force. As a result, and as the USAF and DoD continues to operate in an increasingly virtual world, the increase in simulated training environments must be met with more intentional efforts to understand both the fidelity of simulator-based environments as well as the impact of simulator-training on the ability to meet training requirements.

Simulated Training Environment Fidelity

All training environments where training is not on-the-job can be classified as simulated training environments, whether they are comprised of live, virtual, constructive, and/or augmented elements. Within such environments, simulated training fidelity is described as the "degree to which the training devices must duplicate the actual equipment" (Allen, 1986). However, both objective (e.g., mathematical) and subjective (e.g., psychological) definitions of fidelity have been used to understand and demonstrate the impact of fidelity on training (Lefor et al., 2020). The primary question concerning training environment fidelity often focuses on whether or not the simulation represents the real-world well enough to facilitate adequate and accurate learning to support positive transfer without introducing negative training in the form of inaccuracies, "sim-isms," or counterproductive cognitive or physical habits (Hamstra, et al., 2014; Roberts, et.al., 2020).

While there is a substantial body of literature that recognizes the multi-dimensional nature of fidelity, including physical, functional, visual, aural, tactile, kinesthetic, and cognitive/psychological, (Alessi, 2017; Allen, et. al, 1986; Hays, 1980; Rehmann et al., 1995; Schroeder et al., 2014), there lacks a universally accepted terminology to describe simulator-based training fidelity (Stanton, et.al., 2020). Further, in addition to the physical simulator, simulator-based training fidelity can be impacted by aspects such as the training methodology of the simulator user, characteristics of the broader environment the simulator is a part of (e.g., other simulators that are part of the training), and information that is embedded within the simulators (e.g., training scenarios). Despite this, it is evident that the level of fidelity in each dimension that is necessary for effective training is dependent upon the training task (Beaubien & Baker, 2004; Hamstra, et al., 2014). Given that many simulated training environments are designed to train a myriad of tasks, a top-down approach to evaluating the overall fidelity of a simulated training environment is too cumbersome and contextually insensitive. Rather, what is needed is a bottom-up approach that is focused on training requirements and whether or not the simulated training environment can provide adequate levels of fidelity to meet the training requirement threshold.

SIMULATOR FIDELITY ASSESSMENT APPROACH

Background

Within the past decade, the USAF has utilized several simulator assessment and evaluation methods. However, previously, there lacked a systematic approach to understanding the perceptions of training fidelity from the everyday user, or operator. As a result, the following simulator fidelity assessment approach was developed with three explicit goals in mind, all of which specifically addressed key gaps or limitations of existing evaluation methodologies. The first goal was to leverage the experiences of the operator (e.g., the warfighter utilizing simulator-based training) to better understand how simulator training impacts the ability to meet training requirements. More explicitly, the intent of this goal was to utilize the opinions and perceptions of the operators to better understand how simulators facilitate or fail to facilitate adequate training and map that onto to their own established training requirements. The next goal was to expand the diversity of the data to include both qualitative and quantitative data. The intent of this goal was to ensure that the results of the evaluations included data that were standardized and could quantify elements of the simulator training (quantitative date) while also descriptive enough to account for contextual nuances within each simulator (qualitative data). Further, the qualitative data could be used to specifically understand the perceptions of the operators at an individual level. Lastly, the third goal was to create an evaluation that could generate a report quickly and in an accessible format that could help inform leaders when making decisions when addressing inadequate simulator training capabilities. Early prototype stages of this approach were executed using proctored written evaluations to allow for variations in answer responses (e.g., if an operator selected a certain answer response, the operator would be presented with additional questions that corresponded to their selection). To reduce the labor-intensive nature of the evaluations and to modernize the approach, the evaluations were transitioned to a web-based platform which increased the flexibility of the evaluations and minimized the need for face-to-face proctored data collections.

Approach

The following section outlines the simulator fidelity assessment approach using an example from an evaluation within the USAF. This approach, however, is highly flexible and can be altered to assess different domains and types of simulators or training devices. As previously mentioned, this approach is executed using a web-based platform which allows for virtual evaluations and maximum flexibility. See Figure 1 for an overview of the process.

To begin, evaluation content is generated and imported into the web-based platform. Although customizable, the developed evaluation content often leverages established training documentation, such as existing training requirements. For example, previous content has included training tasks from Training Task Lists (TTLs) and Ready Aircrew Program (RAP) Tasking Memos (RTMs). As part of this process, key stakeholders, and subject matter experts (SMEs) are involved in selecting and modifying the evaluation content to ensure that the content is relevant, applicable to simulator training, and appropriately represents training requirements. For example, content from RTMs that refers to training requirements that are performed outside of the simulator would be removed.

Next, a set of primary and secondary deficiencies are developed. Deficiencies are categories and sub-categories that are used to identify technological areas that causes an inability to receive satisfactory training (see Figure 2). Unique to the current approach, deficiencies were utilized in order to facilitate the identification of trends within the data, such as primary and secondary sources of inadequate fidelity. Primary deficiencies represent the overarching category (e.g., Scenarios) while secondary deficiencies represent nested, corresponding sub-categories (e.g., Realism). Similar to the evaluation content, primary and secondary deficiencies are developed in partnership with key stakeholders and SMEs in order to ensure relevance.

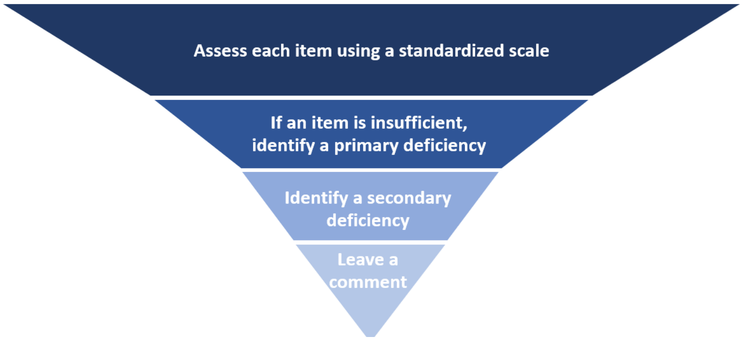
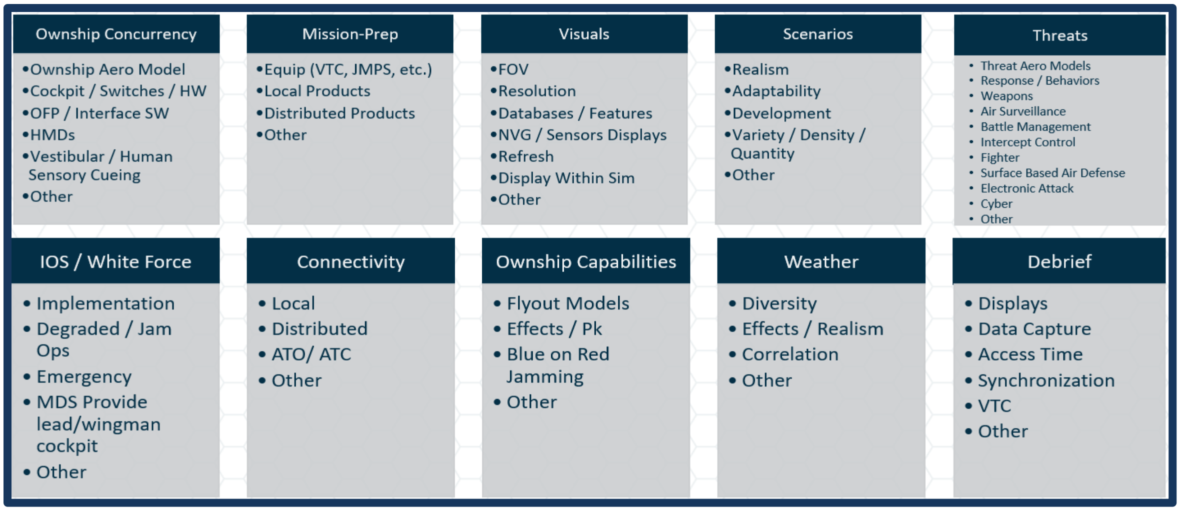

Assess each item using a standardized scale
If an item is insufficient, identify a primary deficiency
Identify a secondary deficiency
Leave a comment

Figure 1. Evaluation Process Overview



Ownship Concurrency

- Ownship Aero Model
- Cockpit / Switches / HW
- OFP / Interface SW
- HMDs
- Vestibular / Human Sensory Cueing
- Other

Mission-Prep

- Equip (VTC, JMPS, etc.)
- Local Products
- Distributed Products
- Other

Visuals

- FOV
- Resolution
- Databases / Features
- NVG / Sensors Displays
- Refresh
- Display Within Sim
- Other

Scenarios

- Realism
- Adaptability
- Development
- Variety / Density / Quantity
- Other

Threats

- Threat Aero Models
- Response / Behaviors
- Weapons
- Air Surveillance
- Battle Management
- Intercept Control
- Fighter
- Surface Based Air Defense
- Electronic Attack
- Cyber
- Other

IOS / White Force

- Implementation
- Degraded / Jam Ops
- Emergency
- MDS Provide lead/wingman cockpit
- Other

Connectivity

- Local
- Distributed
- ATO/ ATC
- Other

Ownship Capabilities

- Flyout Models
- Effects / Pk
- Blue on Red Jamming
- Other

Weather

- Diversity
- Effects / Realism
- Correlation
- Other

Debrief

- Displays
- Data Capture
- Access Time
- Synchronization
- VTC
- Other

Figure 2. Example Deficiency Set

Once the evaluation content and deficiencies have been finalized and imported within the platform, the simulator evaluation can occur. The simulator fidelity assessment approach utilizes a drill-down methodology in order to collect both standardized quantitative as well as more flexible and contextual qualitative data. Operators are asked to rate each evaluation item on a Likert scale. Although customizable, previous evaluation response options include "very poor," "poor," "marginal," "good," and "very good." Operators are also given the response options "does not exist" for training capabilities that do not exist within their simulator as well as "did not evaluate" for training capabilities that they are unable to evaluate. For example, an operator may select did not evaluate in the event that they have not experienced the simulator capability despite the fact that it exists within the simulator. If an item is rated as insufficient ("marginal," "poor," or "very poor"), operators are instructed to identify a primary deficiency (e.g., Visuals) as well as one secondary deficiency (e.g., Field of View [FOV]) that best reflects the simulator deficiencies for that training item. In addition, for items rated insufficient, operators are then required to provide an open-ended comment detailing additional information on the deficiency they selected as well as the reason the training item was rated as insufficient.

The results of the evaluation are then exported as a JavaScript Object Notation (JSON) file and uploaded to an Excel report template. The evaluation report contains eight essential pieces of information: (1) a high-level overview of the number of items in each rating category, (2) the highest and lowest-rated items, (3) items that demonstrated both a high degree of agreement and disagreement among raters (interrater reliability calculated using Fleiss' Kappa; Landis & Koch, 1977), (4) a summary of primary and secondary deficiencies, (5) potential training gaps based on training tasks that were not rated, (6) participant demographics, (7) detailed item-by-item summary statistics for each item including rating distribution, inter-rater reliability, numeric average, percentage of 'not rated', and the top primary deficiency; and (8) full rater comments sorted by training category. In sum, the various parts of the report indicate the quality of the training environment using multiple types of data as well as includes statistical principles that examine the quality of the data. Taken together, the information within the report helps to facilitate decision-making across the entire spectrum regarding what can be trained and what cannot; what shortcomings are the biggest problem and whether they can be improved, and which training environments offer the greatest return on investment for training dollars spent.

SIMULATOR EVALUATION

Methods

The following evaluation examined the simulator training fidelity of a Command and Control (C2) platform. Participants in the evaluation included 28 Air National Guard (ANG) operators across two evaluation groups. The sample included both enlisted servicemembers (*n*=24) and officers (*n*=4). The average number of years in service was 10.5 years and the average number of years with the evaluated platform was 8.3 years. On average, the sample had completed 2.5 virtual flag exercises with the evaluated platform. Evaluation group one (*n*=17) consisted of two positions and evaluation group two (*n*=11) consisted of three positions. The two evaluation groups were created based on the positions of the operators and their required training task requirements such that positions with similar training task requirements were aggregated within the same evaluation group. This process was done utilizing input from both SMEs and key stakeholders.

Content for the evaluation was generated from the TTLs and RTMs. For this particular evaluation, there was an emphasis on better understanding simulator training fidelity across different training methodologies, such as the use of a distributed mission operations (DMO) network to conduct simulator training. As a result, the evaluation was structured such that operators evaluated both their local, standalone simulator training as well as their distributed simulator training for each evaluation content item. Given the emphasis on better understanding distributed simulator training, the evaluation also included two additional modules that had multiple choice and open-ended questions regarding distributed simulator training. Questions for these two modules were developed in partnership with key stakeholders and SMEs. Although the training for the sampled unit occurs on the network facilitated by the Distributed Trained Operations Center (DTOC), distributed training was referred to as DMO as specificized by stakeholders due to operator's familiarity with the term as a reference to distributed training.

The evaluation occurred in person as part of a larger training exercise. Operators participating in the evaluation were provided a brief overview of the evaluation process and then registered within the web-based platform. In order to ensure both privacy and security, the operators were able to register with their common access card (CAC) or their email and cellphone, which included a two-factor authentication (2FA) process to access the platform. Operators were provided a laptop and completed the evaluation across three designated time-blocks. The evaluation was conducted in partnership with Air Combat Command (ACC) and the Air Force Research Laboratory (AFRL).

Results

Results from the standalone evaluation demonstrate simulator fidelity gaps for several training items for both evaluation group one and two (see Table 1 for a summary). For evaluation group one, four items had a mean score of marginal. The highest rated item was "Crew Integration", and the lowest rated item was "Link 16 Ops." The average item reliability was .58 (moderate agreement) with three items having poor or lower reliability and nine items having marginal reliability. When looking at the individual items, 38 of the 41 items had at least one deficiency. On average, operators cited 5.94 deficiencies when completing the evaluation. The most common primary deficiency was Scenarios (39 citations), and the most common secondary deficiency was Scenario Realism (23 citations). At least one operator selected "did not evaluate" or "does not exist" for 24 of the items. In total, there were 99 comments for evaluation group one. For example, "System has no capability to simulate electronic attack/jamming on sensor operations."

For evaluation group two, one item had a score of poor or lower and 13 items had a mean score of marginal. The highest rated item was "Air Track ID", and the lowest rated item was "Large Force Employment (LFE) Control." The average item reliability was .48 (moderate agreement) with 10 items having poor or lower reliability and 7 items having marginal reliability. When looking at the individual items, 37 of the 42 items had at least one deficiency. On average, operators cited 9.00 deficiencies when completing the evaluation. The most common primary deficiency was IOS/White Force Exercise Administration (42 citations), and the most common secondary deficiency was Instructor/Operator Station (IOS) Implementation (34 citations). At least one operator selected "did not evaluate" or "does not exist" for 13 of the items. In total, there were 96 comments for evaluation group two. For example, "This is limited by available stations and manning to successfully recreate a realistic scenario. Additionally, there is not a dedicated training for the white force, this is just basically sit and execute with no formal training."

Results from the DMO evaluation demonstrated similar results to the standalone evaluation when examining top and bottom rated items as well as number of deficiencies for evaluation group one and two, albeit with some key differences. For evaluation group one, three items had a mean score of marginal. The highest rated item was "Crew Integration", and the lowest rated item was "Config Sys Console." The average item reliability was .61 (substantial agreement) with only one item having poor or lower reliability and seven items having marginal reliability. Similar to the standalone evaluation, 38 of the 41 items had at least one deficiency. On average, operators cited 5.35 deficiencies when completing the evaluation. The most common primary deficiency was Connectivity (38 citations), and the most common secondary deficiency was CNT Distributed (23 citations). At least one operator selected "did not evaluate" or "does not exist" for 22 of the items. In total, there were 84 comments for evaluation group one. For example, "Non-chat availability creates a bottleneck in the amount of interfaces between a crew and the DMO personnel."

Lastly, for evaluation group two, seven items had a mean score of marginal. The highest rated item was “Transmit Tactical Reference Point (TRP)”, and the lowest rated item was “Perform CC Ops.” The average item reliability was .58 (moderate agreement) with seven items having poor or lower reliability and six items having marginal reliability. When looking at the individual items, 34 of the 42 items had at least one deficiency. On average, operators cited 7.18 deficiencies when completing the evaluation. The most common primary deficiency was Scenarios (47 citations), and the most common secondary deficiency was Scenario Realism (26 citations). At least one operator selected “did not evaluate” or “does not exist” for 13 of the items. In total, there were 96 comments for evaluation group two. For example. “DMO's are usually better than local SIMs though it's not always the case. I've been in DMO scenarios where we control off a phone-like line versus the utilization of radios.”

Table 1. Results Summary Table

Evaluation Group	Highest Rated Item	Lowest Rated Item	# of Insufficient Items	Avg. Reliability	Avg. Cited Deficiencies	Top Primary Deficiency	Top Secondary Deficiency	# of Comments
Group 1: Standalone	Crew Integration	Link 16 Ops	4	.58	8.42	Scenarios	Scenario Realism	99
Group 2: Standalone	Air Track ID	LFE Control	14	.48	12.38	IOS/White Force	IOS Implementation	96
Group 1: DMO	Crew Integration	Config Sys Console	3	.61	9.10	Connectivity	CNT Distributed	84
Group 2: DMO	Transmit TRP	Perform CC Ops	13	.58	9.88	Scenarios	Scenario Realism	96

In addition to the individual evaluation reports, a comparative report was generated to better examine differences in standalone and DMO training ratings (See Figure 3). Analyses revealed few differences between the training environments. Descriptively, DMO items were rated slightly higher compared to standalone simulator training. For evaluation group one DMO was rated better in 44% of the items compared, Standalone was rated better in 27%, and the remaining 29% had identical ratings. The average rating per item was 3.91 for DMO and 3.87 for Standalone. For evaluation group two, DMO was rated better in 76% of the items compared to just 14% for Standalone and 10% of the items were rated the same. The average rating per item was 3.82 for DMO and 3.58 for Standalone. Paired t-tests were completed for individual items and overall average ratings for each group and revealed no statistical differences. Additional analyses were conducted to determine if the difference between ratings could be attributed to experience or position, but ANCOVAs revealed that neither experience nor position were significant covariates, although statistical power was very low given very small *ns* for individual positions.

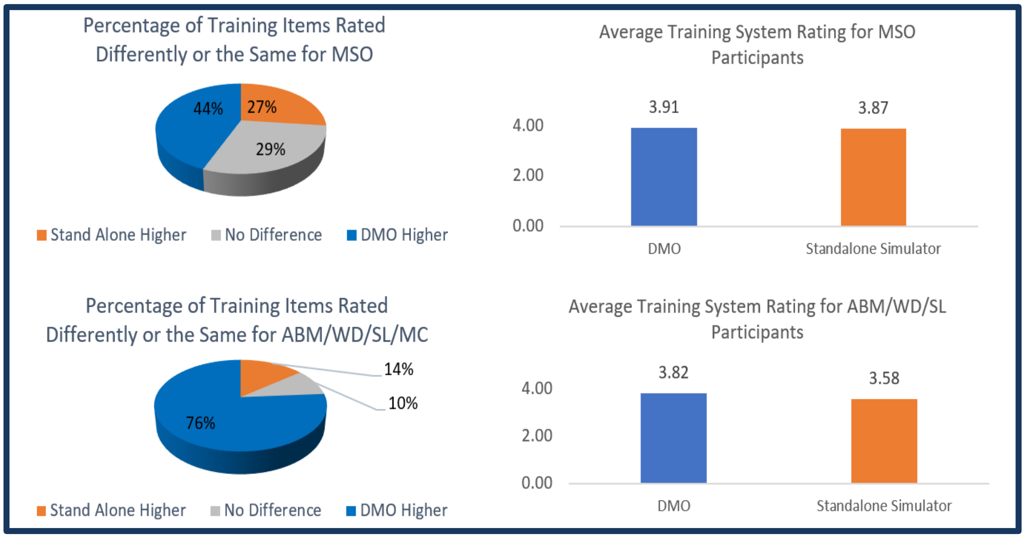


Figure 3. DMO and Standalone Results

Key Themes and Limitations

Results from the evaluation demonstrate both strengths of simulator training as well as areas of improvement for both standalone and DMO training. Although several simulator training deficiencies were highlighted, a few key themes emerged. To begin, responses for both Standalone and DMO training highlighted limitations of training scenarios. For the standalone simulator training, comments frequently cited the need to generate local scenarios in-house with minimal specialized support. For DMO training, comments frequently cited the limited number of scenarios available and the lack of both variety within the scenarios as well as complexity. Specifically, comments noted a lack of scenarios with advanced adversaries and 5th generation platforms. Another key theme included limitations to mission playback and debrief capabilities. For standalone simulator training, comments reflected a lack of on-site debrief capabilities. For DMO training, comments cited limited playback capabilities from the DTOC. Lastly, and one of the more frequently cited comments, regarded limitations to simulator degraded operations training. For both standalone and DMO training, comments cited that communications and equipment degraded operations training are lacking. For example, operators noted that simulated radio communications provide low fidelity training when compared to live experiences with radio communications such that within live experiences, radio communications are often not as clear as simulated radio communications.

There are several limitations that may impact the results of the evaluation. To begin, although nearly the entirety of a unit was sampled, the sample is small compared to the larger community that represents the evaluated platform. In addition, the sample comes from an ANG unit, which may have unique differences in training, equipment, and experience when compared to an active-duty unit. Both of which may have impacted the strength of the relationship when examining differences between DMO and standalone simulator training ratings. More explicitly, although the data demonstrated increased ratings for DMO, the sample size and nuances within the sample size (e.g., experience level) may have contributed to the lack of significant findings. As a result, an effort is underway to collect additional data for the evaluated platform, to include data from an active-duty unit.

BEST PRACTICES

As previously mentioned, the described approach represents one potential method of assessing simulator training fidelity. However, when assessing simulator training fidelity or simulated training environments more broadly, there are several best practices that should be considered that are not specific to the aforementioned approach.

To begin, it is crucial that feedback and simulator fidelity information is collected, and that how that data are collected is carefully considered. Although this may seem like a fundamental and obvious best practice, oftentimes "fidelity assessments" are based on a few individual's experiences through the use of informal, ad hoc inputs from trainees, operators, and evaluators. Instead, simulator fidelity assessments should seek to implement a standardized process or approach to gathering the data with specific questions or points of understanding in mind. This will facilitate the collection of data as well as ensure that the data collected are usable and can accurately and appropriately address gaps in simulator fidelity.

In addition, it's important to consider simulator fidelity wholistically. Within the USAF, simulator fidelity is made up of several parts such as the physical simulator equipment, the scenarios and threats displayed within the simulator, and the cross-platform training engagements that occur as part of simulator training. In any given situation, one part may be disproportionately and adversely impacting the perception of fidelity. For example, an operator may perceive the simulator equipment as providing insufficient training when in actuality, the scenarios being developed for the simulator training are contributing to a lack of fidelity and training transfer. Further, trainees, operators, and evaluators may not always understand or know where problems occur that impact simulator training fidelity. As a result, it's important to consider all parts of simulator training fidelity when determining both areas of adequate fidelity and areas of improvement.

Next, it is important to consider both individual responses as well as the data at an aggregated level. For example, individual responses, especially qualitative responses, often demonstrate individual differences in simulator experiences. More explicitly, individual qualitative responses may illuminate more nuanced simulator perceptions that can be the result of individualized experiences such as inadequate training or lack of experience. On the other hand, individual qualitative responses can also demonstrate a more robust and descriptive understanding of problems with a simulator given repeated exposure over time. Similarly, individual responses can also provide key contextual insights. For example, it may be the case that more experienced operators are rating a content item, on average, as providing sufficient training while less experienced operators are rating the same item, on average, as insufficient. Examining the individual qualitative responses can provide contextual information that explains the difference (e.g., difference in training, difference in expectations when it comes to technology, etc.). Although individual qualitative responses can be incredibly insightful, it's important to also consider the data as a whole. Data at an aggregated level often more accurately reflects broad issues, opposed to one person's individual response. When making key decisions such as where to spend time, money, and effort when addressing simulator fidelity, data at the aggregate-level are more likely to provide a reliable reflection of widespread, critical issues, especially in larger samples. In many cases, the loudest or "squeakiest" voices can shape beliefs about the quality or fidelity of simulator training. As a result, it is key to have data at an aggregate-level that demonstrates the ground truth.

Lastly, while simulator fidelity assessments are targeted at better understanding the trainer, not the trainee, simulator fidelity can also provide key insights regarding training outcomes. For example, in the described approach, potential training gaps were highlighted using a "did not evaluate" and "does not exist" response option. This option reflected training items that operators had not experienced because they did not have the opportunity to experience the item or weren't sure whether the simulator has or will have the capability. Given that the purpose of simulator fidelity assessment is often not to evaluate training outcomes, such findings may require follow-up evaluations or focus groups to better understand training gaps.

CONCLUSION

In conclusion, simulator fidelity assessments are critical for understanding simulated training environment strengths and areas of improvement. The current paper presents a simulator fidelity assessment approach that leverages the perceptions of the operator in order to better understand simulator fidelity in the context of training requirements. Results from an evaluation demonstrate critical areas of improvement that are limiting perceptions of adequate training as a result of simulator deficiencies. Further, results demonstrate trends in rating differences between standalone simulator training and distributed training that need to be further explored, including addressing limitations of the current effort. Taken together, simulators and simulated training environments must continue to be assessed to understand the extent to which they are providing training that allows for establishing training requirements to be met. Simulator assessments are essential to allocate resources and training hours to those training environments that provide the greatest return on investment – balancing the monetary and human costs and risks with the associated projected gains in readiness. Additionally, assessments reveal other essentials such as user buy-in, practical training gaps, force-wide training gaps, and diagnostic information to direct training environment improvements. This is especially important as simulators and other virtual trainers become more popular and as the USAF and the military at large continues to lean into a digital training landscape that provides varying levels of training fidelity.

ACKNOWLEDGEMENTS

The authors would like to acknowledge ACC and AFRL for facilitating and supporting this effort as well as the operators who participated in the evaluation process.

REFERENCES

Alessi, S. (2017). Simulation design for training and assessment. *In Simulation in Aviation Training* (pp. 47-72). Routledge.

Allen, J.A., 1986, *Maintenance training simulator fidelity and individual difference in transfer of training, Hum. Factors, 28*(5), 497–509.

Beaubien, J. M., & Baker, D. P. (2004). *The use of simulation for training teamwork skills in health care: how low can you go?. BMJ Quality & Safety, 13*(suppl 1), i51-i

Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B., & Cook, D. A. (2014). *Reconsidering fidelity in simulation-based training. Academic medicine, 89*(3), 387-392.

Hays, R. T. (1980). *Simulator Fidelity*. U.S. Army Research Institute for the Behavioral and Social Sciences.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lefor, A. K., Harada, K., Kawahira, H., & Mitsuishi, M. (2020). The effect of simulator fidelity on procedure skill training: a literature review. *International journal of medical education, 11*, 97–106. https://doi.org/10.5116/ijme.5ea6.ae73

Liu, D., Macchiarella, N.D., & Vincezi, D.A. (2008). Simulation Fidelity. Hancock, P.A., Vincenzi, D.A., Wise, J.A., Mouloua, M. (Eds.). *Human Factors in Simulation and Training* (1st ed.). CRC Press. https://doi.org/10.1201/9781420072846

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

Ragan, E. D., Bowman, D. A., Kopper, R., Stinson, C., Scerbo, S., & McMahan, R. P. (2015). Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task. *IEEE Transactions on Visualization and Computer Graphics, 21*(7), 794–807. https://doi.org/10.1109/tvcg.2015.2403312

Rehmann, A. J., Mitman, R. D., Reynolds, M. C., & Center, T. (1995). *A Handbook of Flight Simulation Fidelity Requirements for Human Factors Research*. Federal Aviation Administration Technical Center.

Roberts, A. P., Stanton, N. A., Plant, K. L., Fay, D. T., & Pope, K. A. (2020). You say it is physical, I say it is functional; let us call the whole thing off! Simulation: an application divided by lack of common language. *Theoretical issues in ergonomics science, 21*(5), 507-536.

Schroeder, M., Schreiber, B. T., Freiman, M., & Kegley, J. (2014). *Training Effectiveness of the FA-18 Tactical Operational Flight Trainer (TOFT) Upgraded with a Motion-Cueing Seat and Improved Visual System*. Public Release.

Spencer, B. (2009). *The Precious Sortie: The United States Air Force at the Intersection of Rising Energy Prices, an Aging Fleet, a Struggling Recapitalization Effort, and Stressed Defense Budgets.* https://apps.dtic.mil/sti/pdfs/ADA516463.pdf