Fidelity Trade-Offs for Deployable Training and Rehearsal

Brian T. Schreiber
Lumir Research Institute
Tempe, AZ
Brian.Schreiber@mesa.afmcaf.mil

Winston Bennett, Jr.
Air Force Research Laboratory
Mesa, AZ
Winston.Bennett@mesa.afmc.af.mil

Sara Elizabeth Gehr
The Boeing Company
Mesa, AZ
Liz.Gehr@mesa.afmc.af.mil

ABSTRACT

Distributed Mission Operations (DMO) training frequently relies on simulation to accomplish training objectives. Fidelity, in its broader definition and interpretation, encompasses both physical attributes (e.g., ergonomics, switches, symbols, etc.) and functional attributes (e.g., dynamics, models, exercising appropriate cognitive skills, etc.). On a continuum of possible "fidelity levels" what degree of physical and functional fidelity, however, constitutes "high fidelity" or "low fidelity?" What standard is employed and measured against to assign such labels? And, most importantly, what training trade-offs exist when sacrificing higher fidelity for lower cost? That is, in efforts to lower costs (and therefore fidelity), what training experiences are most sacrificed and how is that documented? In this paper we outline a method for evaluating simulation fidelity based upon a comprehensive list of warfighter-defined experiences critical to performing his/her job—that is, a proposed approach upon which simulation systems could be judged and compared. This proposed warfighter-centric approach leverages two credible processes/products already in existence, the Mission Essential Competencies (MECs) and Dash One Emergency Procedures (EPs). During the MEC process, operational warfighters determine the critical list of mission experiences necessary to be fully prepared for combat, while the Dash One lists the critical EPs with which a warfighter must be familiar. Leveraging and combining these products provides us with a warfighter-anchored approach for what constitutes critical simulator system evaluation points. Using this method, we also report the trade-off results between a mature "high-fidelity" and an early deployable "low-fidelity" F-16 four-ship DMO environment. We discuss how the fidelity levels clearly impact each system's ability to provide training on various tactical and EP experiences.

ABOUT THE AUTHORS

Brian T. Schreiber is CEO and Senior Scientist with Lumir Research Institute in support of the Air Force Research Laboratory, Warfighter Readiness Research Division, in Mesa, AZ. He completed his M.S. in Human Factors Engineering at the University of Illinois at Champaign-Urbana in 1995.

Winston Bennett, Jr. is a Senior Research Psychologist and team leader for the training systems technology and performance assessment at the Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division, in, Mesa AZ. He received his Ph.D. in Industrial/Organizational Psychology from Texas A&M University in 1995.

Sara Elizabeth Gehr is a Human Factors Design Specialist with the Boeing Company. She works at the Air Force Research Laboratory, Warfighter Readiness Research Division, in Mesa, AZ, where she is the task order lead for coalition and international mission training research. She received her Ph.D. in Experimental Psychology from Washington University in St. Louis in 2001.

Fidelity Trade-Offs for Deployable Training and Rehearsal

Brian T. Schreiber
Lumir Research Institute
Tempe, AZ
Brian.Schreiber@lumirresearch.com

Winston Bennett, Jr.
Air Force Research Laboratory
Mesa, AZ
Winston.Bennett@mesa.afmc.af.mil

Sara Elizabeth Gehr
The Boeing Company
Mesa, AZ
Liz.Gehr@mesa.afmc.af.mil

INTRODUCTION

The use of simulators to train pilots has become more and more of a necessity with the increase in restrictions on (and high cost of) live-fly exercises (Schank, Thie, Graf, Beel, & Sollinger, 2002). In addition, the ability to network simulators together to allow pilots to train as they fight – as a team – has increased the possible uses for simulators in training. Although numerous studies have shown the effectiveness of simulator training (Bell & Waag 1998; Gehr, Schreiber, & Bennett, 2004; Schreiber & Bennett, 2006), the design and level of functionality of these simulators vary widely. They may range from a desktop computer with a rudimentary stick and throttle and generic weapons, to a motion, full field of view replication of the actual cockpit and specific weapons loads. These varying levels of fidelity both functional (the dynamics of the plane and missiles) and physical (the look and feel of the system, Allen, Buffardi, & Hays, 1991) may affect what can be trained in the simulators (e.g. basic flight maneuvers versus advanced tactics), and the amount of training benefit the pilot will receive.

Although this range of fidelity currently exists in simulator systems, there has been no structured, standardized evaluation process of what level of fidelity is necessary to train different mission experiences. This is a complicated question, and the first step is to derive a comprehensive list of experiences that a warfighter must be able to perform. These experiences must cover all possible aspects of a pilot's potential mission, from routine to the unexpected.

Mission Essential Competencies (MECs^{SM})

To generate the list of all possible experiences, we wanted to start with a list of combat experiences in which every pilot should be competent before being combat ready. The Mission Essential Competency process is one possible way to get a comprehensive list of necessary experiences. MECs are "higher-order individual team, and inter-team competencies that a fully prepared pilot, crew, flight, operator, or team requires for successful mission completion under adverse conditions and in a non-permissive environment" (Colegrove & Alliger, 2002). As part of the MEC process, subject matter experts (SMEs) generate a list of mission *experiences* in which a pilot must be competent before being fully prepared for a combat mission. The MECs have been generated for a great number of domains and platforms (e.g., F-15, A-10, Weapons Directors, Forward Air Controllers, etc.). A very useful application of the MEC experiences is to identity gaps in training. That is, what important MEC experiences are not currently being trained in a given environment? For a study of fidelity issues in training for any given domain/platform, these MEC experiences can provide a common warfighter-defined basis for comparing simulators with varying levels of fidelity.

Procedural Skills: Emergency Procedures

Using simulators to train higher level combat experiences is a relatively recent development in the history of simulator training. The MEC experiences, by their very nature of defined combat-relevant mission experiences, assume the warfighter has already reached a minimum proficiency with lower-level procedural skills/knowledge. The procedural skill areas of greatest relevance to aviation—the domain of interest in the current work—are Emergency Procedures. Before simulators were able to be networked together, they were used in a stand alone mode to train pilots on these basic, procedural elements. This aspect of simulator procedure training still continues to be done today. Previous research has shown that even low fidelity simulators can be used to train inexperienced pilots in basic procedures (Dennis & Harris, 1998). Thus, in order to have comprehensive system evaluation, evaluation of simulator fidelity should include procedural elements in which pilots must be proficient in addition to the experiences enumerated by the MEC process.

Emergency procedures are one example of procedural information with which a pilot must be familiar. We chose to use emergency procedures as a low level procedure because pilots' ability to train on those tasks in a simulator may depend on the fidelity of the simulator. The emergency procedures for F-16 pilots are outlined in their aircraft systems and procedures manual, referred to as the Dash One. One chapter of this book covers the procedures a pilot must follow for almost any emergency situation the pilot will encounter. These bold face critical action emergency procedures (those required to be committed to memory) are used in conjunction with the MECs as the basis of the current study.

Current Work

The goal of the current study is to use the list of MEC experiences and the emergency procedures as the basis of a study of fidelity trade-offs. This list of diverse mission and emergency pilot experiences can be used to compare simulators with different levels of fidelity. Or, similarly, the MEC experiences and EPs could be used to evaluate the same system longitudinally in order to evaluate the training value gained with increased fidelity investment. In the current work, we began with an initial “low-fidelity” F-16 Deployable Tactical Trainer (DTT) and desired to understand its training limitations when compared against more advanced (i.e., what most would label “high fidelity”) stand-alone F-16 simulation systems. This is also a very low-cost approach to assess simulator fidelity needs, without the need to spend additional, possibly unnecessary, funds. We will also present preliminary results of the comparison, based on the list of experiences, by SMEs, of two existing simulator systems, a high fidelity four-ship (referred to hereafter as the DART system), and the lower fidelity deployable four-ship system (hereafter referred to as the DTT system).

METHOD

Evaluators

Eight F-16 subject matter experts independently evaluated the DTT, while six raters independently evaluated the high-fidelity DART system. All evaluators were given advance copies of the experiences to be rated, detailed documentation of the system components, capabilities, and limitations, and were provided multiple opportunities to fly various missions in the system before they gave ratings on the system were given.

The eight DTT evaluators were all male, all recently retired (3.8 years) F-16 pilots from the United States Air Force (USAF), and had an average number of operational F-16 flight hours of 1,949 (average total flight hours of 3,562).

The six DART evaluators were all male, all recently retired (3.9 years) from USAF, and had an average number of operational F-16 flight hours of 1,763 (average total flight hours of 3,500).

Survey

The evaluation survey consisted of 198 total items to be rated: 72 air-to-ground (A/G) MEC experiences, 55 suppression of enemy air defense (SEAD) MEC experiences, 44 air-to-air (A/A) MEC experiences, and the 27 bold face EPs (critical action procedures required to be committed to memory) for the F-16. The scale for each item to be rated was:

0= N/A. Capability to experience does not exist.
1= Capability to experience exists, but is very poor.
2= Capability to experience exists, but is poor.
3= Capability to experience exists, but is marginal
4= Capability to experience exists, and is good
5= Capability to experience exists, and is very good

Apparatus: “Low fidelity” DTT system

The DTT system evaluated consisted of the following major hardware/software functionalities (see Figure 1):

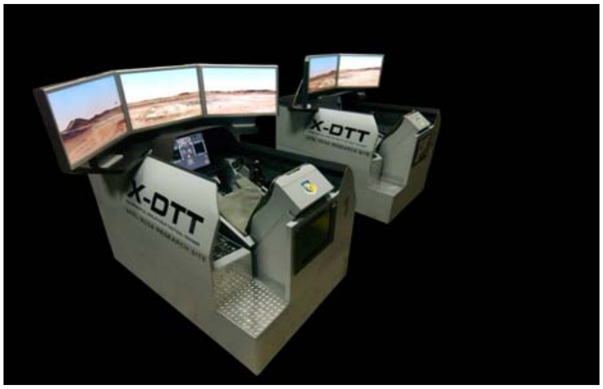

Figure 1. Two DTT cockpits (brief/debrief hardware not shown).

Hardware: F-16 cockpit shell with three out-the-window 30-inch displays, the actual F-16 stick/throttle, but only simulation of essential cockpit switches on a touch screen. Image generator was an SDS International AAcuity® PC-IG system. Brief/debrief includes SmartBoard and two 50-inch displays for mission playback of the Heads-Up Display, Radar Warning Receiver, and both Multi-Function Displays.

Software: Avionics from F-16 Block 30 (SCU3), three databases available, unclassified threat generation system, debrief software with ability to link & time-synchronize video recordings from multiple players, and the ability to network through Distributed Interactive Simulation (DIS) or High Level Architecture (HLA) standards. Only generic, unclassified weapons are available.

Apparatus: "High fidelity" DART system

Hardware: The high-fidelity stand-alone simulators (Figure 2) have all the same hardware elements of the DTT, but are improved in the following key general areas. The high-fidelity simulators are configured with actual F-16 aircraft subpanels and switches – all of which are functional. The out-the-window view consists of an eight-channel, eight-window, state-of-the-art, rear-projection, photo-realistic visual display system providing approximately 20/60 visual acuity. It is a full-color, high-resolution, wraparound display system designed for use with single-seat cockpit simulators. The brief/debrief utilizes six 50-inch screens, dedicating more display real estate to aircraft avionics and providing a replay of the weapon director's tactical display.

Software: The high fidelity simulators have additional fidelity in the following key general areas. The hardware runs classified Block 30 (SCU 5p) actual operational F-16 software. Three different classified threat generation systems are available with high fidelity aerodynamic and weapons modeling. F-16 weapons are all higher fidelity classified models, and include additional weapons such as AIM-120, high energy laser, JDAM, CBU-87, etc. Electronic emissions are modeled as are electronic and infrared countermeasures. Emissions are displayed on an ALR-69 Radar Warning Receiver and chaff/flare is controlled by an ALQ-213 CMDS unit.

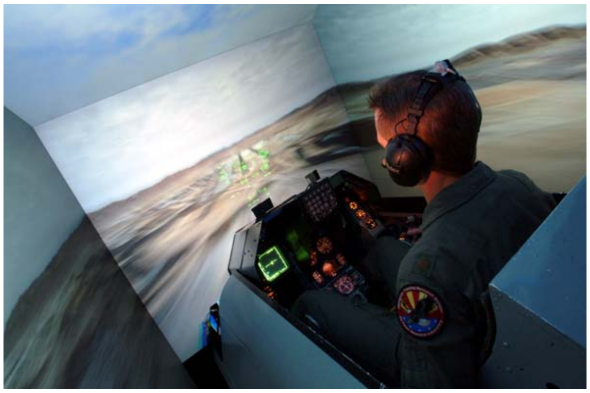

Figure 2. "High fidelity" DART simulator.

RESULTS

The average rating, standard error, rater agreement, and inter-rater reliabilities were calculated. Results for the DTT system are shown in Table 1. While reliability calculations were performed using common correlation procedures (Glass, G. V. & Hopkins, K. D., 1996), the agreement calculations were made after converting the rating responses into a discrete decision choice of either N/A (i.e., 0 ratings) or some degree of applicability (i.e., any rating 1 or higher).

Over all items, the average rating for the DTT system was 1.26, with the subcategory of A/A rated highest (1.79) and the Emergency Procedures subcategory lowest (.59). Raters of the "low fidelity" DTT system across all items unanimously rated 36 of the 198 items as a 0, and 81% of all ratings were between 0 and 2.

Table 1. DTT evaluation results

	avg	s.e.	agreement	reliability
A/G	1.46	.12	.82	.50
A/A	1.79	.14	.83	.53
SEAD	.91	.12	.80	.56
EPs	.59	.04	.47	.18
All items	1.26	.07	.77	.54

The average rating, standard error, rater agreement, and inter-rater reliabilities were also calculated for the DART system, results for which are shown in Table 2. Over all items, the average rating for the DART system was 1.71. Similar to the DTT results, the DART results also revealed the mission subcategory of A/A rated highest (2.56) and the Emergency Procedures subcategory lowest (.67). Raters of the DART system across all items unanimously rated 28 of the 198 items as a 0, and 53% of all ratings were between 0 and 2.

Table 2. DART evaluation results

	avg	s.e.	agreement	reliability
A/G	2.16	.15	.83	.64
A/A	2.56	.20	.84	.75
SEAD	.96	.13	.71	.60
EPs	.67	.08	.57	.47
All items	1.71	.10	.77	.66

The average rating across all items for the DART system was significantly greater than the average rating across all items for the DTT system, t(394)=3.80, p<.05. The average ratings for the A/G items and the A/A items were significantly greater for the DART system, with t-values of 3.64 and 3.18 respectively (both p<.05). The average rating for the SEAD and EP items were not significantly different between the two systems. The comparison of the DTT and DART systems by mission/experience subcategory--A/G, A/A, SEAD, EPs--is presented Figures 3-6, respectively. Independent sample t-tests were calculated across all 198 experiences. 53 of the 198 experiences were rated significantly higher (t-statistics for these ranging between 2.58 and 23.25, all p's<.05) for the DART system than the DTT system. No experiences were rated significantly higher for the DTT over the DART. Further examination of the 53 significant experience differences revealed that 26 were A/A experiences, 26 were A/G experiences, 0 were SEAD experiences, and 1 was an EP.

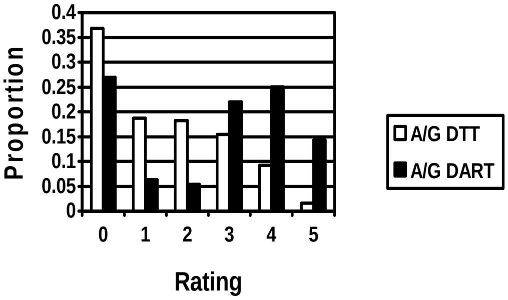

Figure 3. DTT/DART system comparison: Histogram of proportion of responses for A/G experiences

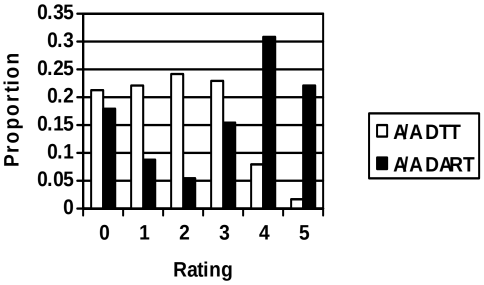

Figure 4. DTT/DART system comparison: Histogram of proportion of responses for A/A experiences

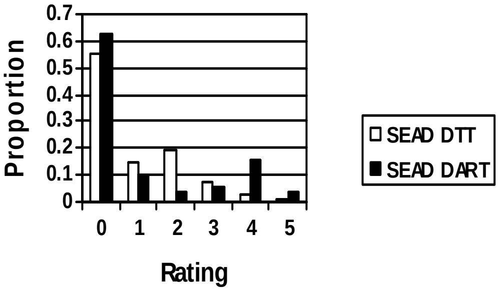

Figure 5. DTT/DART system comparison: Histogram of proportion of responses for SEAD experiences

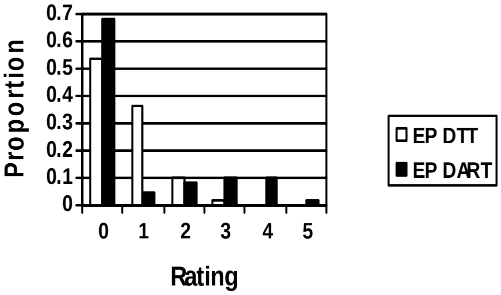

Figure 6. DTT/DART system comparison: Histogram of proportion of responses for emergency procedures

DISCUSSION

Using MEC experiences and Dash One EPs is a different system evaluation approach. As a different approach, a primary concern going into our first set of evaluations was validity of the entire experience survey rating system/process. In the current study, as expected, the DART system overall was rated much higher than the DTT system. Providing additional face validity, specific experiences were rated as one would expect given that the systems are simulators. For example, the experiences of "G-induced physical limitations" and "Live weapons employment (e.g.

WSEP, combat)" were both rated as 0, or N/A. Adding additional validity, save for the DTT EPs, agreements and reliability ratings were sufficiently high. Furthermore, when examining specific differences, 53 of the 198 experiences were significantly different, and **every** one of those 53 differences showed the DART system more highly rated than the DTT system. These results provide us with the validity necessary to not only interpret meaningful differences between the DTT and DART systems, but also to extend this same system evaluation process to other systems. That is, to leverage both the experiences from the MEC process and formal procedural tasks outlined in existing military documents as formal system evaluation points.

The results here clearly show that the DART system is a better system overall. Though the DART system overall was rated higher, it is nonetheless noteworthy that both systems were rated fairly low overall (1.26 and 1.71). We attribute this finding to (a) some functions do not exist in either system (e.g., Joint Stand-Off Weapon, laser targeting and designation, High Speed Anti-Radiation Missile, Mavericks), and (b) a potential rater bias towards the lower end of the scale. Support for the potential low rating scale bias argument comes from another recently completed study (Schreiber, Silverberg-Rowe, & Bennett, 2006), where the same DART system evaluated here was also rated by 32 operational F-16 pilots. In that study, pilots rated the DART as higher than all six other types of training environments for providing the A/A MEC experiences. The DART system results were statistically no different than the highly valued **live-fly** training events such as Flag exercises. Though the results very clearly indicated the DART system as possessing very high training utility, the authors found that the overall DART rating was still just 2.65 on a 0-4 scale (4 being highest).

Interestingly, though A/A and A/G were rated as the highest mission experience areas for the DTT, those were the same (and only) aggregated mission areas where the DART system was rated significantly higher. We attribute the fact that A/A and A/G were highest rated due to development of both these systems with a priority on A/A and A/G capability (as opposed to tasks such as EPs), hence the highest ratings for each system were in these two mission areas. Of course, the primary difference between the systems was the fidelity (and cost) investment—the DART system receiving additional fidelity in technological areas that greatly benefit A/A and A/G experiences.

Follow-up interviews with the evaluators indicated that the underlying reasons for the A/A and A/G differences between the two systems were due to increased investment in the DART for technologies such as (a) near 360 degree high resolution visuals, (b) classified friendly missile models, and (c) classified threat models, all three of which impact the ability of pilots to accurately and correctly execute A/A and A/G tactics. Indeed, when examining the 53 individual experiences found to be statistically different, many of those A/A and A/G MEC experiences rely heavily on one or more of the above mentioned technologies. A few examples of these experiences include "Full range of adversary ground type and mix…", "Simulated Weapons Employment", "Formation responsibilities (e.g., position, visual lookout)", and "Lost mutual support." These last two examples provide, for purposes of a paper, a simple and concise explanation as to why these were significantly different rated experiences between the DTT and DART systems: Flying various formations cannot be adequately experienced/performed in a system without a full 360 degree visual field.

As the DART system contains the full compliment of actual cockpit switches while the DTT system does not, we expected to find a number of significant differences for the EPs (but did not). The evaluators, however, were not at all surprised by the non-significant EP results. Though the DART system has the capability to activate the various switch positions (and have subsequent effects on the aircraft systems), there is currently no mechanism to inject the necessary fault. That is, fault insertion from the instructor operator station (IOS) which would lead to the required cockpit switch actuation procedure(s) from the pilot, is not possible. Evaluators were quick to point out that making this one simple technological change in the IOS would greatly increase many of the DART EP ratings.

In an effort to dovetail the evaluation process results identified here into usable, concrete system development, the next step in our research process will be to identify the systematic technological deficiencies underpinning many of the low DTT/DART ratings (e.g., the IOS fault insertion). Those deficiencies will then be placed in a matrix survey against the same 198 experiences. All matrix cells will be filled in with ratings of the extent to which each technological deficiency detracts from the ability to gain each experience. Technological deficiencies can then be rank ordered in terms of their pervasiveness within and across the different mission areas. The end goal of this subsequent step will be to (a) provide data-driven recommendations for system improvements, and (b) maximize the dollars spent on improvements that will yield the greatest increased value for training utility, either within or across the different mission areas. In other words, this second step to the research process can identify the engineering and procurement requirements to efficiently and effectively upgrade a simulation system.

ACKNOWLEDGEMENTS

The authors wish to thank the USAF Air Combat Command for sponsoring this research and development work. This research was conducted under USAF Contract No. F41624-97D-5000. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

Allen, J., Buffardi, L., & Hays, R. (1991). *The Relationship of Simulator Fidelity to Task and Performance Variables* (ADA238941). Fairfax, VA: George Mason University..

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology*. 8 (3), 223-242.

Colegrove, C. M. & Alliger, G. M. (2002, April). *Mission Essential Competencies: Defining Combat Mission Readiness in a Novel Way*. Paper presented at the NATO RTO Studies, Analysis and Simulation Panel (SAS) Symposium. Brussels, Belgium.

Dennis, K. A. & Harris, D. (1998). Computer-based simulation as an adjunct to ab initio flight training. *International Journal of Aviation Psychology*, 8 (3), 261-276.

Gehr, S. E., Schreiber, B. T., & Bennett, W. Jr. (2004). Within-Simulator Training Effectiveness Evaluation. In *2004 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Proceedings*. Orlando, FL: National Security Industrial Association.

Glass, G. V. & Hopkins, K. D. (1995). *Statistical Methods in Education and Psychology*, 3rd Ed. Needham Heights, MA: Allyn and Bacon.

Schank, J. F., Thie, H. J, Graf, C. M., Beel, J., & Sollinger, J. (2002). *Finding the right balance: Simulator and Live Training for Navy Units*. Santa Monica: RAND Corporation.

Schreiber, B. T. & Bennett, W. Jr.(in press). *Distributed mission operations within-simulator training effectiveness baseline study. Summary report.* (AFRL-HE-AZ-TR-2006-0015-Vol I). Air Force Research Laboratory, AZ: Warfighter Readiness Research Division.

Schreiber, B. T, Rowe, L., & Bennett, W. Jr. (in press). *Distributed mission operations Within-simulator training effectiveness baseline study. Participant utility and effectiveness opinions and ratings.* (AFRL-HE-AZ-TR-0015-Vol IV). Air Force Research Laboratory, AZ: Warfighter Readiness Research Division.