Fidelity requirements for effective training: Pilot perceptions versus objective results

Ms. Jamie L. Estock, Ms. Kathryn Baughman, Dr. Emily M. Stelzer, Dr. Amy L. Alexander
Aptima, Inc.
Woburn, MA
jestock@aptima.com, kbaughman@aptima.com, estelzer@aptima.com, aalexander@aptima.com

ABSTRACT

Fidelity requirements defined by users provide valuable insight into the fidelity needed to ensure that trainees 'buy-in' to the simulator as a training device. However, there are no empirical data to support a relationship between trainees' perceptions of a simulator's training effectiveness and actual training effectiveness. Our preliminary research revealed a discrepancy between pilots' perceptions of the effectiveness of the simulator as a training device and objective *in-simulator performance* results (Estock, Alexander, Stelzer, & Baughman, 2007). For this paper, we conducted additional analyses to determine whether a similar discrepancy exists between pilots' perceptions of training effectiveness and objective *training effectiveness* results. Specifically, we conducted an experiment in which 43 U.S. Air Force F-16 pilots flew air-to-air training research missions. During the experimental trials, two pilots flew in high-fidelity F-16 simulators with a 360° field of view (FOV), and two pilots flew in lower-fidelity F-16 simulators with a 108° FOV. Both before and after these experimental trials, all pilots flew benchmark missions using only the high-fidelity simulator. To obtain objective assessments of the *training effectiveness* of each simulator, we compared the two groups on their change in performance on air-to-air skills *from pre- to post-training benchmark missions*. To obtain subjective assessments of the training effectiveness of each simulator, we administered a questionnaire to all pilots immediately following the experimental trials. We focused on the effectiveness of each simulator in training a set of air-to-air skills most likely to be influenced by the FOV differences between the two simulators. We compared trainees' perceptions of training effectiveness with objective training effectiveness results. The findings of this study replicated the findings of our previous study in that we found a discrepancy between pilots' perceptions and objective results. We discuss the implications of these findings for the verification, validation, and accreditation (VV&A) of training simulators.

ABOUT THE AUTHORS

Ms. Jamie Estock is a Human Factors Scientist and Team Lead in the Human Performance Division at Aptima, Inc. Ms. Estock leads Aptima's line of work focused on identifying fidelity requirements to support effective training and developing decision-support tools to provide guidance for purchasing and employing training systems. Ms. Estock holds a M.A. in Human Factors/Applied Cognitive Psychology from George Mason University and a B.S. in Psychology from the University of Pittsburgh at Johnstown.

Ms. Kathryn Baughman is an Industrial/Organizational Analyst in the Human Performance Division at Aptima, Inc. Ms. Baughman has experience investigating the effects of training environments (e.g., simulators, games) on training effectiveness and developing evaluations for cognitive skill-based training programs. Ms. Baughman is currently a doctoral candidate in Industrial-Organizational Psychology at George Mason University. She received an M.A. in Industrial-Organizational Psychology from George Mason University and a B.S. in Psychology from the University of Georgia.

Dr. Emily M. Stelzer is a Cognitive Scientist and Team Lead in the Cognitive Systems Engineering Division at Aptima, Inc. Dr. Stelzer has experience investigating the effects of complex displays and automated systems on human performance. Dr. Stelzer holds a Ph.D. and M.A. in Engineering Psychology from the University of Illinois at Urbana-Champaign, and a B.A. in Psychology from the University of Cincinnati.

Dr. Amy L. Alexander is a Human Factors Scientist and Team Lead Scientist in the Human Performance Division at Aptima, Inc. Dr. Alexander has experience evaluating advanced flight deck technologies and assessing human performance in complex environments. Dr. Alexander holds a Ph.D. and M.A. in Engineering Psychology from the University of Illinois at Urbana-Champaign, and a B.S. in Psychology from The Ohio State University.

Fidelity requirements for effective training: Pilot perceptions versus objective results

Ms. Jamie L. Estock, Ms. Kathryn Baughman, Dr. Emily M. Stelzer, Dr. Amy L. Alexander
Aptima, Inc.
Woburn, MA
jestock@aptima.com, kbaughman@aptima.com, estelzer@aptima.com, aalexander@aptima.com

INTRODUCTION

Simulator fidelity requirements defined by users provide valuable insight into the fidelity needed to ensure that trainees 'buy-in' to a simulator as a training device. If trainees do not perceive the simulator to be an effective training device, they may be less motivated to fully engage in the training. Since trainee motivation has been linked to training effectiveness (e.g., Mumford, Weeks, Harding, & Fleishman, 1988; Noe, 1986), we would expect this lack of buy-in to result in poor training effectiveness.

However, there are no empirical data to support a relationship between trainees' perceptions of a simulator's effectiveness and actual training effectiveness results. In fact, our preliminary research revealed a discrepancy between pilots' perceptions of the effectiveness of the simulator as a training device and objective *in-simulator performance* results (Estock, Alexander, Stelzer, & Baughman, 2007). To obtain objective assessments of *in-simulator performance*, we compared the performance of trainees *while flying the training missions* in two simulators of differing levels of fidelity. The current paper investigates whether a similar discrepancy exists between pilots' perceptions of the effectiveness of the simulator as a training device and objective *training effectiveness* results. To obtain objective assessments of *training effectiveness*, we compared trainees on their change in performance *from pre- to post-training benchmark missions* after flying training missions in two simulators of differing levels of fidelity.

In our experiment, F-16 pilots flew air-to-air training missions in two different simulators—the high-fidelity Display for Advanced Research and Technology (DART) simulators and the lower-fidelity Deployable Tactics Trainer (DTT) simulators—at the Air Force Research Laboratory in Mesa, Arizona (AFRL/Mesa). The primary difference between the DART simulators and the DTT simulators is the size of the horizontal visual scene field-of-view (FOV). The high-fidelity DART simulators have a 360-degree horizontal FOV visual system whereas the lower-fidelity DTT simulators have a 108-degree horizontal FOV visual system. A narrow FOV can provide a keyhole view of the world, which may limit awareness of peripheral regions of the visual scene (Wickens, Thomas, & Young, 2000; Woods, 1984). However, a simulator with a narrow FOV may only have a negative impact on the ability to train those air-to-air skills most affected by FOV, while proving to be an effective device for training other air-to-air skills.

The skills pilots must acquire to be considered mission-ready for air-to-air combat were identified through the Mission Essential Competencies (MECs[SM]) process (Colegrove & Alliger, 2002). A survey distributed to ten F-16 subject matter experts (SMEs) prior to this experiment suggested that the reduced FOV in the lower-fidelity DTT could negatively impact the effectiveness of the simulator in training two skills that are largely dependent on visual information: (1) the ability to maintain a briefed formation, and (2) the ability to defeat or deny the threat in a visual arena. Specifically, 100 percent of F-16 SMEs surveyed reported that the ability to train the skill of *maintaining formation* would be affected by the FOV of the simulator. In addition, 100 percent of the F-16 SMEs surveyed reported that the ability to train the skill of *executing merge gameplan* would be affected by the FOV of the simulator. As a result, we expect that pilots flying the lower-fidelity DTT simulators will report lower subjective ratings of the effectiveness of the simulator at training the skills of *maintaining formation* and *executing merge gameplan* than pilots flying the high-fidelity DART simulators. However, based on our previous research, we expect to find a discrepancy between pilots' perceptions of the effectiveness of the simulator as a training device and objective training effectiveness results.

METHODS

Participants

Forty-three U.S. Air Force F-16 pilots participated in the experiment. All 43 participants were male. The majority of participants (58%) held the rank of Captain, and the majority of participants (58%) were Instructor Pilots. The participants had a mean of five years flying the F-16 aircraft (*SD* = 2.64). The participants had a mean of 958 total F-16 hours (*SD* = 617.72), and a mean of 77 F-16 hours in the past six months (*SD* = 35.33).

Simulators

The DART simulators are high-fidelity simulators consisting of an F-16 Block 30 aircraft cockpit and running the F-16 Block 30 aircraft's Operational Flight Program (OFP). The DART simulators contain the actual F-16 aircraft controls and displays. The simulators have a 360-degree horizontal FOV visual system with 1600x1200 pixel resolution display. Figure 1 provides a view from inside the DART simulator.

Figure 1. DART simulator.

The DTT simulators are lower-fidelity simulators consisting of an F-16 Block 30 aircraft 'shell' and running the F-16 Block 30 aircraft's OFP. The DTT simulators use a high-fidelity aircraft stick and throttle, and have the essential F-16 cockpit switches on a touch screen LCD in front of the pilot. The DTT simulators have a 108-degree horizontal FOV visual system with 2560x1600 pixel resolution display. Figure 2 provides a view of the DTT simulator.

Figure 2. DTT simulator.

Experimental Design

The experimental design contrasted pilots who flew the lower-fidelity DTT simulators with pilots who flew high-fidelity DART simulators on: (1) perceptions of the effectiveness of the simulator as a training device, and (2) objective training effectiveness results. The design focused on the impact of visual scene FOV on the effectiveness of the simulator at training two air-to-air skills—*maintaining formation* and *executing merge gameplan.*

The design compared the subjective evaluation of the effectiveness of the simulator as a training device between pilots flying the lower-fidelity DTT simulators versus the high-fidelity DART simulators. The effectiveness ratings were obtained post-training via questionnaire. Specifically, the pilots were asked to rate their level of agreement with the following statements—using a one to five Likert scale from strongly disagree to strongly agree: (1) The simulator was an effective way to train me how to *maintain a briefed formation*, and (2) The simulator was an effective way to train me how to *defeat or deny the threat in a visual arena.*

The design also contrasted the objective training effectiveness results between pilots flying the lower-fidelity DTT simulators versus the high-fidelity DART simulators. The training effectiveness results were captured via a comparison of the change in performance on pre- and post-training benchmark missions between pilots flying the lower-fidelity DTT simulators versus the high-fidelity DART simulators. We used the Performance Evaluation

Tracking System (PETS; Schreiber, Watz, Bennett, & Portrey, 2003) to obtain objective, simulator-based performance data associated with the skills of *maintaining formation* and *executing merge gameplan.*

The F-16 SMEs identified four PETS measures that assess a pilot's ability to *maintain briefed formation*. All four measures were taken when the pilots were within 40 nautical miles (NM) of an enemy. These measures included: (1) the average two-dimensional (2D) range between flight lead and wingman, (2) the average three-dimensional (3D) range between flight lead and wingman, (3) the number of mutual support violations, and (4) the total time spent in mutual support violation. A mutual support violation occurs when a pilot is outside of normal formation parameters and is less able to support his flight lead/wingman with his weapons or radar.

The F-16 SMEs identified two PETS measures that assess a pilot's ability to *execute merge gameplan*. Specifically, these measures included: (1) the number of viper mortalities, and (2) the number of enemy kills.

Procedures

The four-day training research experiment at AFRL/Mesa consisted of two experimental sessions per day. The participants flew standard air-to-air missions from AFRL/Mesa's Distributed Mission Operations (DMO) Training Research Syllabus as an integrated team of four (a "four-ship"). Prior to flying the training missions, the four-ship flew three benchmark missions with all four pilots flying in the high-fidelity DART simulators. Then, we randomly assigned the pilots to either the high-fidelity DART or lower-fidelity DTT condition for 17 training missions. During the training missions, two pilots flew in the high-fidelity DART simulators and two pilots flew in the lower-fidelity DTT simulators. After the final training mission, the pilots completed a questionnaire to evaluate the effectiveness of the simulator as a training device. Finally, the four-ship flew three additional benchmark missions with all four pilots flying in the high-fidelity DART simulators. During the pre- and post-training benchmark missions, objective performance data were collected in real-time via PETS.

RESULTS AND DISCUSSION

Independent samples *t* tests were used to determine if there were statistically significant differences between: (1) the pilots' subjective evaluation of the effectiveness of the lower-fidelity DTT simulators and the high-fidelity DART simulators, and (2) the training effectiveness of the lower-fidelity DTT simulators and the high-fidelity DART simulators. A p value of ≤ 0.05 was considered a statistically significant difference. Effect size was also calculated to determine whether a statistically significant difference was also practically significant. We used Cohen's d as our measure of effect size where small effect: $d = 0.2$; medium effect: $d = 0.5$; and large effect: $d = 0.8$.

Training effectiveness was measured by change in performance on PETS measures from pre- to post-training on three equivalent benchmark missions. Because there are limitations on how PETS data may be reported (i.e., means cannot be reported in the public domain), we report percent difference in performance from pre- to post-training for both simulator conditions. However, we conducted independent samples *t* tests on the mean change scores between simulator conditions.

Maintain Formation

Pilot Perceptions

The pilots' mean ratings of the simulator's effectiveness for training pilots to *maintain formation* are presented in Figure 3.

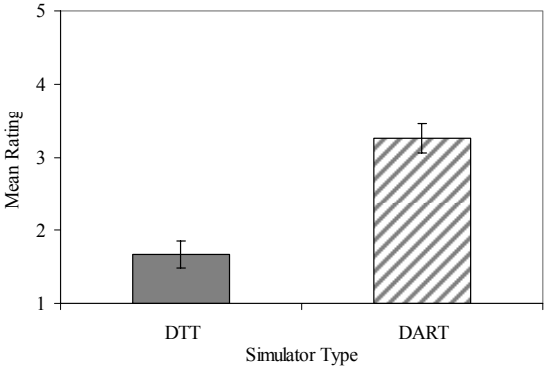

Figure 3. Mean ratings of simulator's ability to train pilots to *maintain formation* by simulator type.

An independent samples *t* test revealed that pilots who flew the lower-fidelity DTT simulators rated the simulator as significantly less effective for training the skill of *maintaining formation* ($M = 1.67$) than pilots who flew the high-fidelity DART simulators

(M = 3.26, $t(41)$ = 4.26, $p < 0.001$, d = 1.24). Furthermore, the large effect size of 1.24 indicates the practical significance of the finding.

Training Effectiveness Results

Table 1 presents percent differences in pre- to post-training performance on PETS measures related to *maintaining formation* for pilots who flew the lower-fidelity DTT simulators and pilots who flew the high-fidelity DART simulators.

Table 1. Pre- to post-training percent difference in *maintaining formation* between pilots flying DTTs versus pilots flying DARTs.

PETS measure	% Difference Post-Training DTT	% Difference Post-Training DART	*t* statistic	*p* value	*d* value
Average 2D Range	-40.90%	107.68%	0.94	0.36	0.29
Average 3D Range	-52.19%	101.96%	1.06	0.30	0.33
Number of Mutual Support Violations*	13.45%	-28.85%	-2.07	0.05	0.88
Total time in Mutual Support Violation*	-1.46%	5.94%	0.49	0.63	0.18

* Decrease (i.e., negative value) is an indicator of improved performance

Independent samples *t* tests revealed no significant difference in average 2D range, average 3D range, or the total time spent in mutual support violation across simulator conditions.

However, an independent samples *t* test on the number of mutual support violations revealed that pilots who flew the lower-fidelity DTT simulators showed less improvement on the number of mutual support violations post-training compared to pilots who flew the high-fidelity DART simulators ($t(41)$ = -2.07, p = 0.05, d = 0.88). Furthermore, the large effect size of 0.88 indicates the practical significance of this training effectiveness finding.

These findings showed that pilots rated the lower-fidelity DTT simulators less effective at training the skill of *maintaining formation* than the high-fidelity DART simulators. However, the training effectiveness results showed no difference in training effectiveness between pilots who flew the lower-fidelity DTT simulators as compared to pilots who flew the high-fidelity DART simulators on three out of four of the objective performance measures related to *maintaining formation*.

Execute Merge Gameplan

Pilot Perceptions

The pilots' mean ratings of the simulator's effectiveness for training pilots to *execute merge gameplan* are presented in Figure 4.

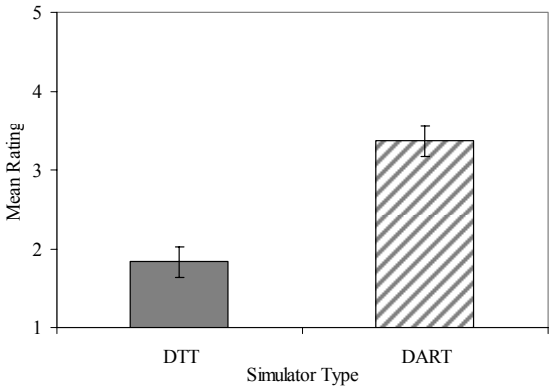

Figure 4. Mean ratings of simulator's ability to train pilots to *execute merge gameplan* by simulator type.

An independent samples *t* test revealed that pilots who flew the lower-fidelity DTT simulators rated the simulator significantly less effective for training the skill of *executing merge gameplan* (M = 1.83) than pilots who flew the high-fidelity DART simulators (M = 3.37, $t(41)$ = 3.95, $p < 0.001$, d = 1.21). Furthermore, the large effect size of 1.21 indicates the practical significance of the finding.

Training Effectiveness Results

Table 2 presents percent differences in pre- to post-training performance on measures related to *executing merge gameplan* for pilots who flew the lower-fidelity DTT simulators and pilots who flew the high-fidelity DART simulators.

Table 2. Pre-to post-training percent difference in *executing merge gameplan* between pilots flying lower-fidelity DTTs versus pilots flying DARTs.

PETS measure	% Difference Post-Training DTT	% Difference Post-Training DART	*t* statistic	*p* value	*d* value
Number of Mortalities*	-64.79%	-74.60%	-0.06	0.95	-0.01
Number of Enemy Kills	4.51%	10.08%	0.35	0.73	0.10

* Decrease (i.e., negative value) is an indicator of improved performance

Independent samples *t* tests revealed no significant difference in the number of mortalities, or the number of enemy kills across simulator conditions.

These findings showed that pilots rated the lower-fidelity DTT simulators less effective at training the skill of *executing merge gameplan* than the high-fidelity DART simulators. However, the training effectiveness results showed no difference in training effectiveness between pilots who flew the lower-fidelity DTT simulators as compared to pilots who flew the high fidelity DART simulators on all objective performance measures related to *executing merge gameplan*.

CONCLUSIONS

We examined the influence of visual scene FOV differences on: (1) pilots' perceptions of a simulator's effectiveness at training two air-to-air skills, and (2) objective training effectiveness results. The air-to-air skills selected for this investigation were *maintaining formation* and *executing merge gameplan*—two skills that are largely dependent on visual information. The results of this study show a discrepancy between pilots' subjective assessment of the effectiveness of simulators and the objective performance outcomes. Specifically, the results revealed that pilots rated the lower-fidelity DTT simulators as less effective for training the ability to *maintain formation* and *execute merge gameplan* than the high-fidelity DART simulators. However, training effectiveness results showed no difference between the lower-fidelity DTT simulators and high-fidelity DART simulators for training the ability to *maintain formation* and *execute merge gameplan*. The training effectiveness results may be due to less than adequate visual fidelity for training the ability to *maintain formation* and *execute merge gameplan* in both the DART and DTT simulators. As a result, the pilots who flew either simulator may have relied on sensor information instead of visual information to perform these skills thereby eliminating the simulator difference most relevant for these skills. Information gathered from post-experiment interviews with pilot SMEs supported this explanation. As a result, future analyses should be conducted to examine if a discrepancy exists between pilots' subjective assessment of the effectiveness of simulators and the objective training effectiveness results for air-to-air skills most affected by cockpit fidelity, such as interpreting sensor output and radar mechanization.

We believe the discrepancy between pilots' subjective assessment of the effectiveness of simulators and the objective training effectiveness results has important implications for verification, validation, and accreditation (VV&A) of training simulators. VV&A are related but distinct processes aimed at gathering and evaluating information to determine, based on the simulation's intended use, a simulators capabilities, limitations, and performance relative to the real-world system it simulates (DMSO, 2002). Some current processes for VV&A involve subjective assessments conducted by users. For example, during the dynamic portion of the U.S. Air Force's Combat Air Forces (CAF) simulator certification (SIMCERT) process, pilot SMEs conduct a detailed subjective evaluation of the simulator. The pilot SMEs rate the simulator's ability to train specific tasks (Chapman, 2006). These subjective assessments are essential to ensure that users buy-in to the simulator as a training device. However, we believe subjective evaluations should be coupled with data collected from objective evaluations to provide the most accurate view of a simulator's effectiveness.

ACKNOWLEDGEMENTS

This material is based upon work supported by the AFRL under Contract No. FA8650-06-C-6649. We would like to thank Dr. Winston Bennett at AFRL/Mesa for funding this research. We would also like to thank Lt. Brenda Blueggel, our technical point of contact, for her support on this effort.

REFERENCES

Chapman (2006). Accreditation Policy and Practice for Immersive Warfighter Simulators. In *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, Florida.

Colegrove, C. M., & Alliger, G. M. (2002). Mission Essential Competencies: Defining Combat Mission Readiness in a Novel Way. Paper presented at: *NATO Research & Technology Organization, Studies, Analysis, and Simulation Panel, Conference on Mission Training via Distributed Simulation* (SAS 38), Brussels, Belgium.

Defense Modeling and Simulation Office (DMSO), (2002) *Recommended Practices Guide (RPG) for Verification Validation and Accreditation (VV&A)*, online, https://www.dmso.mil/public/transition/vva/policiesguidance, 20 June 2006.

Estock, J. L., Alexander, A. L., Stelzer, E. M., & Baughman, K. (2007). Impact of simulator fidelity on F-16 pilot performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 75-79). Santa Monica, CA: HFES.

Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review, 11,* 736-749.

Mumford, M. D., Weeks, J. L., Harding, F. D., & Fleishman, E. A. (1988). Relations between student characteristics, course content, and training outcomes: An integrative modeling effort. *Journal of Applied Psychology, 73,* 443-456.

Schreiber, B. T., Watz, E., Bennett, W. Jr., & Portrey, A. (2003). Development of a distributed mission training automated performance tracking system. In *Proceedings of the 12th Conference on Behavior Representation in Modeling and Simulation,* Scottsdale, AZ.

Wickens, C.D., Thomas, L.C., & Young, R. (2000). Frames of reference for the display of battlefield information: Judgment-display dependencies. *Human Factors*, 42, 660-675.

Woods, D.D. (1984). Visual momentum: A concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies*, 21, 229-244.