Is cockpit fidelity important for effective training? Perception versus performance

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

ABSTRACT

Recent research revealed a discrepancy between pilots' perceptions of training effectiveness and objective training effectiveness results for simulators with different visual scene field-of-view (FOV) sizes. The results indicated that pilots rated the simulator with the narrower FOV less effective for training two air-to-air skills largely dependent on visual information—maintaining formation and executing merge gameplan—than the simulator with the wider FOV, yet the training effectiveness results showed no difference between the two simulators. The purpose of the current research is to examine whether a similar discrepancy exists between pilots' perceptions of training effectiveness and objective training effectiveness results for air-to-air skills *most influenced by cockpit fidelity*. The air-to-air skills selected for this investigation were *interpreting sensor output* and *radar mechanization*—two skills largely dependent on information provided by the cockpit displays. We conducted an experiment in which 43 U.S. Air Force F-16 pilots flew air-to-air missions as an integrated team of four. During the experimental trials, two pilots flew a high-fidelity simulator with the actual F-16 aircraft controls and displays, and two pilots flew a lower-fidelity simulator with the essential F-16 cockpit switches on a touch screen LCD in front of the pilot. Both before and after the experimental trials, all pilots flew three benchmark missions in the high-fidelity simulator. To obtain objective assessments of the training effectiveness of each simulator, we compared pilots who flew the high fidelity simulator to pilots who flew the low fidelity simulator on their change in performance on air-to-air skills from pre- to post-experiment benchmark missions. To obtain subjective assessments of the training effectiveness of each simulator, we administered a self-report questionnaire to all pilots immediately following the experimental trials. In this paper, we compare trainees' perceptions of training effectiveness with objective training effectiveness results and discuss implications for training simulator acquisition and use.

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.

Dr. Emily M. Stelzer is a Cognitive Scientist and Team Lead in the Cognitive Systems Engineering Division at Aptima, Inc. Dr. Stelzer work focuses on 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;s work focuses on 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.

Dr. Kathryn Engel is an Industrial/Organizational Psychologist in the Human Performance Division at Aptima, Inc. Dr. Engel work focuses on investigating the effects of training environments (e.g., simulators, games) on training effectiveness and developing evaluations for cognitive skill-based training programs. Dr. Engel holds a Ph.D. and M.A. in Industrial-Organizational Psychology from George Mason University and a B.S. in Psychology from the University of Georgia.

Is cockpit fidelity important for effective training? Perception versus performance

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

INTRODUCTION

Recent research (Estock, Alexander, Stelzer, & Baughman 2007; Estock, Baughman, Stelzer, & Alexander, 2008) revealed a discrepancy between pilots' perceptions of training effectiveness and objective training effectiveness results for simulators with different visual scene field-of-view (FOV) sizes. The results indicated that pilots rated the simulator with the narrower FOV less effective for training two air-to-air skills largely dependent on visual information—*maintaining formation* and *executing merge gameplan*—than the simulator with the wider FOV, yet the training effectiveness results showed no difference between the two simulator conditions.

The purpose of the research reported in this paper is to examine whether a similar discrepancy exists between pilots' perceptions of training effectiveness and objective training effectiveness results for air-to-air skills *most influenced by cockpit fidelity*. In our experiment, U.S. Air Force F-16 pilots flew air-to-air missions as an integrated team of four. Two pilots flew in high-fidelity Display for Advanced Research and Technology (DART) simulators outfitted with the actual F-16 aircraft controls and displays, and two pilots flew in lower-fidelity Deployable Tactics Trainer (DTT) simulators with the essential F-16 cockpit switches on a touch screen LCD in front of the pilot.

Prior to the experiment, we administered a survey to six F-16 subject matter experts to identify the impact of cockpit fidelity differences on training effectiveness. The F-16 SMEs suggested that cockpit fidelity is the most important fidelity dimension for training mission-ready pilots in air-to-air beyond visual range (BVR) engagements. The pilots engaged hostile aircraft under BVR conditions in the majority of air-to-air missions flown during the experiment. The F-16 SMEs suggested that the lower-fidelity cockpit in the DTT simulators could negatively impact the effectiveness of the simulator in training mission-ready pilots in air-to-air BVR engagements. In addition, the F-16 SMEs suggested that the lower-fidelity cockpit in the DTT simulator may require the pilots to alter some of the processes they would normally execute in the F-16 aircraft. As a result, we expected:

1. Pilots flying the lower-fidelity DTT simulators to report *lower ratings of satisfaction with the cockpit fidelity* of the simulator than pilots flying the high-fidelity DART simulators.

2. Pilots flying the lower-fidelity DTT simulators to report *lower ratings of effectiveness of the cockpit fidelity* of the simulator for training than pilots flying the high-fidelity DART simulators.

The F-16 SMEs also suggested that the lower-fidelity cockpit in the DTT simulators could negatively impact the effectiveness of the simulator in training two skills that are largely dependent on information provided through the cockpit displays: (1) the ability to *correctly translate 2-D sensor output into a 3-D mental model within an appropriate timeframe and determine the appropriate maneuver to gain tactical advantage*, and (2) the ability to *use radar capabilities to effectively locate and track relevant targets*. These skills, which 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). As a result, we expected:

3. Pilots flying the lower-fidelity DTT simulators to report lower subjective ratings of the effectiveness of the simulator at training the skill of *interpreting sensor output* than pilots flying the high-fidelity DART simulators.

4. Pilots flying the lower-fidelity DTT simulators to report lower subjective ratings of the effectiveness of the simulator at training the skill of *radar mechanization* than pilots flying the high-fidelity DART simulators.

However, based on our previous research (i.e., Estock, et al., 2007, Estock et. al., 2008), we expected to find a discrepancy between pilots' perceptions of the

effectiveness of the simulator as a training device and objective training effectiveness results. Specifically, we expected:

5. Pilots to rate the lower-fidelity DTT simulator less effective for training *interpreting sensor output* and *radar mechanization* than the high-fidelity DART simulators, but the objective training effectiveness results to show no difference between the two simulator conditions.

Training effectiveness was assessed by measuring change in pilot performance from pre- to post-training on objective measures related to the skills of *interpreting sensor output* and *radar mechanization.*

Since we expected that the objective performance measures would show no difference between the two simulator conditions, we also measured pilot subjective workload. Subjective workload ratings have been shown to increase when greater mental resources are invested, even when objective performance measures may not reflect this difference (Yeh & Wickens, 1988). Therefore, subjective workload ratings may be a more sensitive measure of the impact of fidelity on training effectiveness. As a result, we expected:

6. Pilots flying the lower-fidelity DTT simulators to report higher levels of subjective workload than pilots flying the high-fidelity DART simulators.

A discrepancy between pilot's perceptions of training effectiveness and training effectiveness results, if identified, has important implications for decisions regarding the acquisition and use of training simulators.

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), 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 have a 360-degree horizontal FOV visual system with 1600x1200 pixel resolution display, and contain the actual F-16 aircraft controls and displays. Figure 1 provides a view of the DART simulator cockpit.

Figure 1. DART simulator cockpit.

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 have a 108-degree horizontal FOV visual system with 2560x1600 pixel resolution display. They use a high-fidelity aircraft stick and throttle, and have the F-16 cockpit panels and switches on a touch screen LCD in front of the pilot. Specifically, the touch screen LCD displays all of the F-16 cockpit panels and switches in the location they would normally be in the jet, but it cannot display all of the panels at once. For example, if the pilot needs to access to the panels on the left or right of his seat, they can call those panels up to be the main display and access the switches that they need. If the pilot presses any of the function buttons, the same menus will appear as they would in the F-16 aircraft and DART simulators. The touch screen LCD provides visual feedback so that when the pilot touches a switch they will see it flip on or off immediately. The touch screen does not provide any tactile feedback. Figure 2 provides a view of the DTT simulator cockpit.

Figure 2. DTT simulator cockpit.

The DART and the DTT simulators differed in both their visual and cockpit fidelity. To account for the specific impact of cockpit fidelity differences between the simulators, we focused our analysis on two skills that are largely dependent on information provided by the cockpit displays—*interprets sensor output* and *radar mechanization*.

Experimental Design

The experimental design compared pilots who flew the lower-fidelity DTT simulators with pilots who flew the high-fidelity DART simulators on: (1) perceptions of the effectiveness of the simulator as a training device, (2) objective training effectiveness results, and (3) subjective workload ratings. The design focused on the impact of cockpit fidelity on the effectiveness of the simulator in training two air-to-air skills—*interpreting sensor output* and *radar mechanization.*

The design compared the *subjective evaluation of the effectiveness of the simulator* as a training device for pilots flying the lower-fidelity DTT simulators versus those flying 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 two statements—using a one to five Likert scale from strongly disagree to strongly agree: (1) I was satisfied with the level of cockpit fidelity of the simulator during this training, and (2) The level of cockpit fidelity of this simulator was effective for this training.

The pilots were also asked to rate their level of agreement with the following statements about the effectiveness of the simulator for training the two air-to-air skills—*interpreting sensor output* and *radar mechanization*—using the same Likert scale: (1) The simulator was an effective way to train me how to *correctly translate 2-D sensor output into a 3-D mental model within appropriate timeframe and determining the appropriate maneuver to gain tactical advantage*, and (2) The simulator was an effective way to train me how to *use radar capabilities to effectively locate and track relevant targets.*

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 Scenario-based Performance Observation Tool for Learning In Team Environments (SPOTLITE; MacMillan, Entin, Morley, & Bennett, in press) to obtain F-16 SME evaluations of the trainees' performance associated with the skills of *interpreting sensor output* and *radar mechanization.* 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 *interpreting sensor output* and *radar mechanization.*In previous work F-16 SMEs identified four SPOTLITE measures that assess the skill of *interpret sensor output*. These measures include: (1) targeting in accordance with standards, (2) recognizing and reacting to a bandit maneuver, (3) shooting in accordance with shot doctrine, and (4) violations of briefed range criteria. The F-16 SMEs also identified four PETS measures that assess the skill of *interpreting sensor output*. These measures include: (1) the number of fratricides, (2) the number of times the pilot allowed a hostile aircraft to fly within Minimum Abort Range (MAR), (3) the number of times the pilot allowed a hostile aircraft to fly within Minimum Out Range (MOR), and (4) the total time the hostile aircraft spent in MAR violation. MAR and MOR measures indicate the extent to which the pilots stay outside of the adversary weapons engagement zone (WEZ; Schreiber, Stock, Bennett, 2006).

The F-16 SMEs identified two SPOTLITE measures that assess the skill of *radar mechanization*. These measures include: (1) appropriateness of radar setup, and (2) targeting in accordance with standards. The F-16 SMEs also identified two PETS measures that assess the skill of *radar mechanization*. These measures include: (1) the number of enemy kills, and (2) the total time spent in MAR violation.

Finally, the design compared the *subjective workload ratings* between pilots flying the lower-fidelity DTT simulators versus the high-fidelity DART simulators. The subjective workload ratings were captured via the NASA Task Load Index (TLX). This subjective workload rating procedure was developed by NASA Ames Research Center (Hart & Staveland, 1988). The NASA TLX is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demand, Physical Demand, Temporal Demand, Own Performance, Effort, and Frustration.

Procedure

The four-day training research experiment at the Air Force Research Laboratory in Mesa, Arizona (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, and the NASA TLX to evaluate the subjective workload ratings of the simulator. 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, and F-16 SME evaluations were collected in real-time via SPOTLITE. Specifically, F-16 SMEs used SPOTLITE to record unsatisfactory performance related to the skills of *interpreting sensor output* and *radar mechanization* for each pilot.

RESULTS AND DISCUSSION

Independent samples *t* tests were used to determine if there were statistically significant differences between:

1. Pilots' subjective evaluation of the effectiveness of the lower-fidelity DTT simulators and the high-fidelity DART simulators,
2. Objective training effectiveness of the lower-fidelity DTT simulators and the high-fidelity DART simulators, and
3. Subjective workload ratings of the lower-fidelity DTT simulators and the high-fidelity DART simulators.

A p value of ≤ 0.05 was considered to be a statistically significant difference. Effect size was also calculated to determine whether a statistically significant difference has some practical significance, and is not just a statistical artifact. We used Cohen's d as a measure of effect size. Cohen (1988) refers to d= 0.20 as a small effect, d = 0.50 as a medium effect; and d >= 0.80 as a large effect.

Training effectiveness was measured by change in performance on SPOTLITE and 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. We conducted independent samples *t* tests on the mean change scores between simulator conditions. For the SPOTLITE data, we report mean change in the number of unsatisfactory performance ratings from pre- to post-training for both simulator conditions.

Cockpit Fidelity

Pilot Perceptions

The pilots' mean ratings of *satisfaction with the cockpit fidelity* of the simulator are presented in Figure 3.

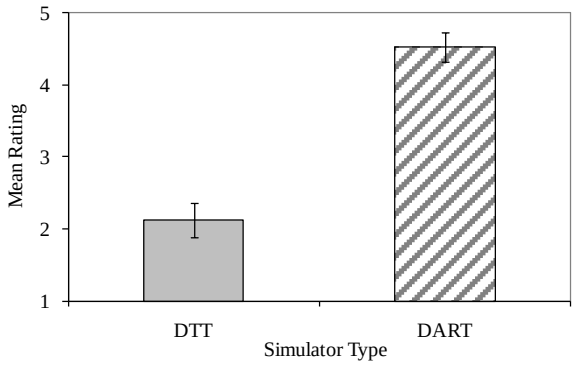

Figure 3. Mean ratings of *satisfaction with the cockpit fidelity* by simulator type.

Independent samples *t* tests revealed that pilots who flew the lower-fidelity DTT simulators reported significantly lower ratings of *satisfaction with the cockpit fidelity* of the simulator (M = 2.12) than the pilots who flew the high-fidelity DART simulators (M = 4.53; $t(41) = -7.28$, $p < 0.001$, d = 2.02). The large effect size of 2.02 indicates the practical significance of the finding.

The pilots' mean ratings of *effectiveness of the cockpit fidelity* of the simulator for training are presented in Figure 4.

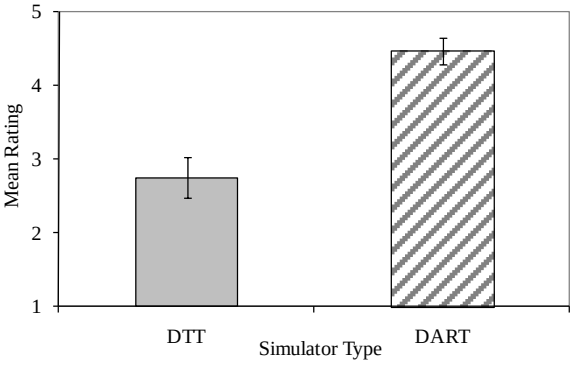

Figure 4. Mean ratings of *effectiveness of the cockpit fidelity* for training by simulator type.

Independent samples *t* tests revealed that pilots who flew the lower-fidelity DTT simulators reported significantly lower ratings of effectiveness of the cockpit fidelity of the simulator for training (M = 2.75) than the pilots who flew the high-fidelity DART simulators (M = 4.47; $t(41) = -5.24$, $p < 0.001$, d= 1.27). The large effect size of 1.27 indicates the practical significance of the finding.

These findings indicate that pilots report lower satisfaction with the cockpit fidelity of the lower-fidelity DTT simulators and rated the cockpit fidelity of this simulator as less effective for training. In the next sections, we examine the extension of this trend to specific skills.

Interprets Sensor Output

Pilot Perceptions

The pilots' mean ratings of the simulator's effectiveness for training pilots to *interpret sensor output* are presented in Figure 5.

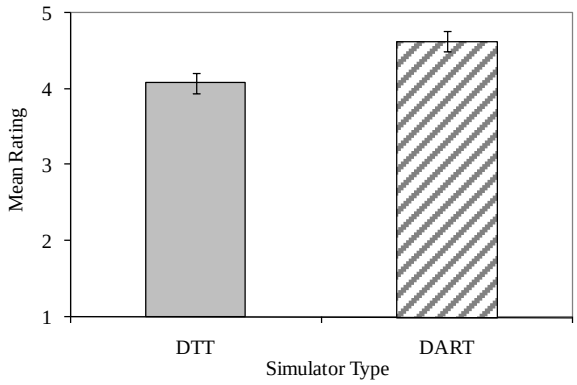

Figure 5. Mean ratings of simulator's ability to train pilots to *interpret sensor output* by simulator type.

Independent samples *t* tests revealed that pilots who flew the lower-fidelity DTT simulators reported significantly lower subjective ratings of the effectiveness of the simulator at training the skill of *interpreting sensor output* (M = 4.08) than pilots who flew the high-fidelity DART simulators (M = 4.63; $t(41) = -2.84$, $p < 0.05$, d= 0.84). The large effect size of 0.84 indicates practical significance of the finding.

Training Effectiveness

Table 1 presents mean differences in pre- to post-training performance on SPOTLITE measures related to *interpreting sensor output* for pilots who flew the lower-fidelity DTT simulators and pilots who flew the high-fidelity DART simulators.

Table 1. Pre- to post-training mean difference on SPOTLITE measures of *interpreting sensor output* between pilots flying DTTs versus pilots flying DARTs.

	Mean Difference Post-Training DTT	Mean Difference Post-Training DART	*t* statistic	*p* value	*d* value
Number of Unsatisfactory SPOTLITE ratings	-1.17	-0.85	0.40	0.69	.02

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

Independent samples *t* tests revealed no significant difference in the number of unsatisfactory performance ratings from pre- to post-training on SPOTLITE measures related to the skill of *interpreting sensor output* across simulator conditions.

Table 2 presents percent differences in pre- to post-training performance on PETS measures related to interpreting sensor output 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 on PETS measures of *interpreting sensor output* between pilots flying DTTs versus pilots flying DARTs.

PETS measure	% Difference Post-Training DTT	% Difference Post-Training DART	*t* statistic	*p* value	*d* value
Number of Fratricides*	0.00%	-100.00%	-1.00	0.33	-0.22
Number of MAR Violations*	-62.80%	-44.24%	1.32	0.19	0.36
Number of MOR Violations*	-9.83%	-10.36%	0.02	0.98	0.01
Total time in MAR Violation*	-74.35%	-60.61%	0.75	0.46	0.21

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

Independent samples *t* tests revealed no significant difference in number of fratricides, the number of MAR violations, the number of MOR violations, or the total time spent in MAR violation from pre- to post-training across simulator conditions.

These findings showed that pilots rated the lower-fidelity DTT simulators less effective at training the skill of *interpreting sensor output* 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 objective performance measures related to *interpreting sensor output*.

Radar Mechanization

Pilot Perceptions

The pilots' mean ratings of the simulator's effectiveness for training pilots on *radar mechanization* are presented in Figure 6.

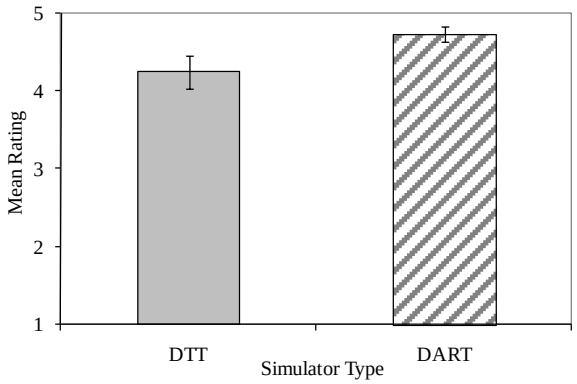

Figure 6. Mean ratings of simulator's ability to train pilots on *radar mechanization* by simulator type.

Independent samples *t* tests revealed that pilots who flew the lower-fidelity DTT simulators (M = 4.25) reported significantly lower subjective ratings of the effectiveness of the simulator at training the skill of *radar mechanization* than pilots who flew the high-fidelity DART simulators (M = 4.74; $t(41) = -2.07$, $p < 0.05$, d= 0.47). Furthermore, the medium effect size of 0.47 suggests that there may be practical significance of the finding.

Training Effectiveness

Table 3 presents mean differences in pre- to post-training performance on SPOTLITE measures related to *radar mechanization* for pilots who flew the lower-fidelity DTT simulators and pilots who flew the high-fidelity DART simulators.

Table 3. Pre- to post-training mean difference on SPOTLITE measures of *radar mechanization* between pilots flying DTTs versus pilots flying DARTs.

	Mean Difference Post-Training DTT	Mean Difference Post-Training DART	*t* statistic	*p* value	*d* value
Number of Unsatisfactory SPOTLITE ratings	-0.57	-0.60	-0.06	0.95	.11

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

Independent samples *t* tests revealed no significant difference in the number of unsatisfactory performance ratings from pre- to post-training on SPOTLITE measures related to the skill of *radar mechanization* across simulator conditions.

Table 4 presents percent differences in pre- to post-training performance on PETS measures related to radar mechanization for pilots who flew the lower-fidelity DTT simulators and pilots who flew the high-fidelity DART simulators.

Table 4. Pre- to post-training percent difference on PETS measures of *radar mechanization* between pilots flying DTTs versus pilots flying DARTs.

PETS measure	% Difference Post-Training DTT	% Difference Post-Training DART	*t* statistic	*p* value	*d* value
Number of Enemy Kills	4.51%	10.08%	0.35	0.73	0.1
Total time in MAR Violation*	-74.35%	-60.61%	0.75	0.46	0.21

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

Independent samples *t* tests revealed no significant difference in number of enemy kills or the total time spent in MAR violation across simulator conditions.

These findings showed that pilots rated the lower-fidelity DTT simulators less effective at training the skill of *radar mechanization* 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 objective performance measures related to *radar mechanization*.

Subjective Workload

Subjective workload ratings have been shown to increase when greater mental resources are invested, even when objective performance measures may not reflect this difference (Yeh & Wickens, 1988). In this study, however, independent samples *t* tests revealed *no significant difference in the overall subjective workload ratings* across simulator conditions.

Although there was no overall workload effect, independent samples *t* tests revealed that pilots who flew the lower-fidelity DTT simulators reported significantly higher subjective ratings on the sub-scale of *frustration* (M = 67.08) than pilots who flew the high-fidelity DART simulators (M = 48.42; $t(41) = 2.48$, $p < .05$, d= 0.70). Furthermore, the medium effect size of 0.70 suggests that there may be practical significance of the finding. The pilots' mean NASA TLX ratings of *frustration* are presented in Figure 7.

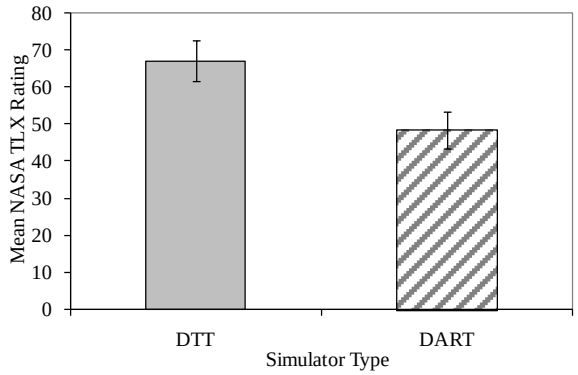

Figure 7. Mean NASA TLX ratings of the *frustration* dimensions by simulator type.

These findings showed no difference in overall workload ratings between pilots who flew the lower-fidelity DTT simulators as compared to pilots who flew the high-fidelity DART simulators. However, the pilots who flew the lower-fidelity DTT simulators reported a higher level of frustration as compared to pilots who flew the high-fidelity DART simulators.

CONCLUSIONS

We examined the influence of cockpit fidelity differences on: (1) pilots' perceptions of a simulator's training effectiveness, (2) objective training effectiveness results, and (3) subjective workload ratings. The air-to-air skills selected for this investigation were *interprets sensor output* and *radar mechanization*— two skills that are largely dependent on information provided by the cockpit displays. The results of this study indicate a discrepancy between pilots' subjective assessment of a simulator's effectiveness and objective performance outcomes.

The subjective results revealed that pilots reported lower satisfaction with the cockpit fidelity of the lower-fidelity DTT simulators and rated the cockpit fidelity in the lower-fidelity DTT simulators less effective for training. An analysis of the two air-to-air skills affected by cockpit fidelity that we assessed revealed that the pilots rated the lower-fidelity DTT simulators as less effective for training the ability to *interprets sensor output* and *radar mechanization* than the high-fidelity DART simulators.

However, objective training effectiveness results showed no difference between the lower-fidelity DTT simulators and high-fidelity DART simulators for training the skills of *interpreting sensor output* and *radar mechanization*.

The SPOTLITE and PETS measures used in this study were identified by F-16 SMEs as meaningful indicators of performance on *interpreting sensor output* and *radar mechanization*. It is possible, though, that the training effectiveness results may be due to the fact that these measures are not at the appropriate level of granularity to detect objective performance differences invoked by varying levels of cockpit fidelity. For example, the performance measures used in this study may not have adequately captured the adjustments that pilots made to account for the lower-fidelity cockpit in the DTT simulators. In post-experiment interviews, pilots who flew the DTT simulators described higher workload resulting from the need to conduct extra steps beyond the normal processes that they would perform in the F-16 aircraft. For example, the touch screen LCD in the DTT cockpit did not display the fuel gauge requiring the pilot to go into the data entry display (DED) for fuel monitoring.

As a result, we investigated the subjective workload rating across simulator condition because previous research has shown a disassociation between performance and subjective measures of workload (Yeh & Wickens, 1988). However, the subjective workload results showed no difference between the lower-fidelity DTT simulators and high-fidelity DART simulators in overall workload. The only significant difference between the lower lower-fidelity DTT simulators and high-fidelity DART simulators was on the NASA-TLX subscale of frustration. This finding suggests that lower-fidelity simulators can cause higher trainee frustration, but that trainee frustration does not manifest into either higher overall workload or reduced training effectiveness.

The discrepancy between pilot's perceptions of training effectiveness and objective training effectiveness results has important implications for decisions regarding the acquisition and use of training simulators. The number and types of technologies available for training pilots is exponentially growing while training in the actual aircraft is becoming increasingly constrained by logistical challenges and limited resources. There is a growing need to ensure that training resources are allocated in a cost-effective manner that optimizes skill acquisition and retention (Ricci, Salas, & Cannon-Bowers, 1996). Those who are responsible for training simulator acquisition and curriculum development need to make difficult decisions regarding which missions and skills can be trained effectively in lower-fidelity simulators, which require higher-fidelity simulators, and which require training in the actual aircraft. Furthermore, they need to make decisions about how to sequence the use of different training devices to maximize training

effectiveness and minimize costs. However, they currently lack objective, concrete measures by which they can determine: (1) what the simulator is capable of training, (2) how to evaluate the simulator's effectiveness, and (3) how to improve future simulators (Estock, McCormack, Bennett, & Patrey, 2008). As a result, pilot SME evaluations drive decisions about which simulators and simulator technologies are acquired and used for training. The pilot SME evaluations are critical to ensure that trainees 'buy-in' to a simulator as a training device and are motivated to fully engage in the training. However, our research demonstrates that the subjective evaluations should be coupled with data collected from objective evaluations to provide the most accurate view of a simulator's effectiveness for training.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second ed.). Lawrence Erlbaum Associates

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.

Estock, J. L., Alexander, A. L., Stelzer, E. M., & Baughman, K. (2007). Impact of visual scene field of view on F-16 pilot performance. Paper presented at the *51st Annual Meeting of the Human Factors and Ergonomic Society*, Baltimore, MD.

Estock, J.L., Baughman, K., Stelzer, E.M., & Alexander, A.L. (2008). Fidelity requirements for effective training: Pilot perceptions versus objective results. *Proceedings of the 30th Annual Interservice/Industry Training, Simulation and Education Conference*, Orlando, FL.

Estock, J. L., McCormack, R., Bennett, W., & Patrey, J. (2008). A Model-based Tool to Quantify Simulation Fidelity: Preliminary Results, Ongoing Development, and Future Applications. Paper presented at the *American Institute of Aeronautics and Astronautics Modeling and Simulation Technologies Conference and Exhibit*, Honolulu, HI.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload.* Amsterdam, The Netherlands: Elsevier.

MacMillan, J., Entin, E. B., Morley, R., & Bennett, W. (in press). Measuring team performance in complex dynamic environments: The SPOTLITE method. Military Psychology.

Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (1996). Do Computer-based Games Facilitate Knowledge Acquisition and Retention? *Military Psychology*, 8, 295-307.

Schreiber, B. T., Stock, W.A., & Bennett, W. Jr. (2006). *Distributed Mission Operations within simulator training effectiveness baseline study: Metric development and objectively quantifying the degree of learning.* (AFRL-HEAZ-TR-2006-0015-Vol II). Mesa, AZ: Air Force Research Laboratory, Warfighter Readiness Research Division.

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.

Yeh, Y-Y., & Wickens, C. D. (1988). The dissociation of subjective measures of mental workload and performance. *Human Factors*, 30, 111-120.