Within-Simulator Training Effectiveness Evaluation

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

Brian Schreiber
Lockheed Martin
Mesa, AZ
Brian.schreiber@mesa.afmc.af.mil

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

ABSTRACT

There exists a need to formally assess the training benefits of Distributed Mission Operations (DMO) training on the performance of F-16 pilots. DMO training consists of multiplayer networked environments designed to enhance warfighter competency. Although many studies have converged on the effectiveness of training pilots in stand alone systems, very little research has been done on the effectiveness of DMO training of multiple pilots in networked simulators. As Bell and Waag (1998) outlined, to establish support for the effectiveness of training, several different levels of converging support are needed. A proper approach would involve collecting data from several sources that taken together will lend support to the significance of DMO training. To establish the effectiveness of networked simulated training, evidence from a variety of sources will be examined, including: (1) objective indicators of the performance of the pilots acting as a four-ship team engaged in point-defense actions, (2) ratings of team performance made by subject matter experts (SME), (3) scaling evidence collected using the Pathfinder paired-comparison methodology, and (4) pilot reactions to DMO as recorded on rating forms collected. Although all four types of data should show support for the effectiveness of DMO training, the inclusion of objective data allows stronger conclusions to be drawn. Objective data enables quantification of the subjective opinions and ratings, thereby providing indications of the return on investment (ROI), in terms of increased human performance, of the training system. Our current work involves assessing pilots using these methods, and the results should address the changes in capability of training our warfighters.

ABOUT THE AUTHORS

Sara Elizabeth Gehr is a Human Factors Design Specialist with Boeing at the Air Force Research Laboratory, Warfighter Readiness Research Division, in Mesa, AZ. She received her Ph.D. in Experimental Psychology from Washington University in St. Louis in 2001.

Brian T. Schreiber is a Staff Scientist with Lockheed Martin at 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.

Within-Simulator Training Effectiveness Evaluation

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

Brian Schreiber
Lockheed Martin
Mesa, AZ
Brian.schreiber@mesa.afmc.af.mil

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

INTRODUCTION

Simulation has historically been a necessary and important part of training pilots. For the more complicated training necessary for combat pilots, simulation has been especially necessary since many aspects of combat flying cannot be trained in-flight for reasons such as peacetime training rules and resource limitations. However, combat simulators have been primarily used to train single pilots on specific aspects of a mission. In contrast, real world missions involve multiple planes and pilots interacting to accomplish a common goal.

Distributed Mission Operations (DMO) Training consists of multiplayer, including networked, environments designed to enhance warfighter competency. That is, multiple pilots in multiple simulators, either at the same physical location or distributed in different locations, can fly and work together on the same mission. In the past several years, a great deal of attention and resources have been lavished on DMO training, the focus of which has generally been on engineering improvements to create a more realistic environment. The questions being addressed typically involve "What" in the simulation environment is not realistic and "How" can we make it more realistic (Watz, Schreiber, Keck, McCall, & Bennett, 2003). However, DMO training environments exist to improve warfighter competence, not simply to create a realistic environment. Investigating and evaluating warfighter competencies as a function of DMO training invites entirely different, non-engineering questions to be addressed, such as which warfighter Mission Essential Competencies (MECs; see Colegrove & Alliger, 2002) benefit from DMO training and, quantitatively, by how much? Though very few doubt DMO training as beneficial, the literature does not provide irrefutable evidence as to the magnitude and types of human performance gains. The literature supports improvement, but the evidence is based solely upon subjective data. A comprehensive DMO training effectiveness evaluation is needed to determine the return on investment, in terms of human performance gains, of a DMO training environment.

Training Effectiveness Evaluations

Bell and Waag (1998) outlined an ideal five stage training effectiveness evaluation model. The first stage is a utility evaluation. This involves gathering feedback from the pilots after they have experienced a training syllabus. This user feedback will provide information about how well the training is received by the pilots, including information about the system fidelity and user acceptance. Without minimum fidelity and acceptance by the pilots, further evaluation would be unnecessary. The second stage of the Bell and Waag evaluation model attempts to measure the amount of learning that takes place within the simulator. A comparison of two similar missions, presented before and after training, will show any effects of the training program. As with the results of the first stage, if no improvement is found within the simulator, additional testing is not necessary. The third stage looks for performance improvements to transfer to a different simulator. That is, the testing environment is a different simulator than the one that was used for the training. Pilots who have had training can be compared to pilots who have not had training when flying the same scenario. The pilots with previous training (even though it was in a different simulator) should perform better than pilots who have not had training. The fourth stage calls for measurement of transfer of training from the simulator to an actual flight environment. However, peacetime flight rules will place limitations on what the pilots can do and resource considerations may reduce the availability of equipment. In addition, it is much harder to control all the variables necessary for a good experimental design in a real-world flight, and harder to collect performance measures. Ideally though, pilots with training would be compared to pilots without training to look for performance improvements. The fifth and final stage of their training effectiveness

evaluation looks at pilot performance in a combat environment. Obviously there are no experimental controls in this study. However, Bell and Waag suggest a modeling approach to look for any transfer of training effects. If the results of all five stages of the training evaluation are significant, this would be convincing evidence of the effectiveness of the training program.

This study focuses on the first and second parts of the model just outlined. Thorough, convincing within-simulator training effectiveness evaluations consist of different types of data converging on the same conclusions. Proper effectiveness evaluations should comprise utility or user opinion data, instructor or rater observation data, and objective data collected both in the simulator and on comparable "real-world" transfer tasks (Bell and Waag, 1998). User opinion data captures what users experience and their opinions on how useful a training system might be, and for which tasks it might be best suited. Instructor or rater observation data provides expert assessment of skill competency change as a function of training. Objective data enables quantification of the subjective opinions and ratings, thereby providing indications of the return on investment (ROI), in terms of increased human performance, of the training system.

Schreiber, Watz, & Bennett (2003) highlighted the need for objective performance assessments to quantify learning improvements in complex networked simulation environments. Additionally, Krusmark & Schreiber (in preparation) identified a number of problematic issues using subjective gradesheet data for explaining systematic performance gains in a DMO environment. Some of the more troublesome issues identified included a potential vested interest by raters to show improvement, an inability to correctly record simple statistics such as kills, and an insufficient observed systematic variance among individually rated skills. Though there was an observed improvement in ratings as a function of DMO, Krusmark & Schreiber concluded that the subjective data alone could not discount the multiple possible explanations for the improvement. The lack of significant systematic variance among a number of rated skills suggests that the subjective measurement system may not be sensitive enough. Furthermore, the observation that objective measures were not counted accurately raised concerns regarding the validity of the other, subjective rating data. Finally, rating scales (as were used in Krusmark & Schreiber) are frequently relative; therefore, absolute quantifiable improvement gains as a function of training are difficult to ascertain.

Using various degrees of opinion, rater, and/or objective data, a fair amount of prior research exists on simulator training effectiveness for simpler tasks representative of a small portion of a mission (e.g., manual bomb delivery, one versus one air combat) and all found simulator training beneficial (Gray and Fuller; 1977; Gray, Chun, Warner, and Eubanks, 1981; Hagin, Dural, and Prophet, 1979; Wiekhorst, 1987; Lintern, Sheppard, Parker, Yates, and Nolan, 1989; Kellogg, Prather, and Castore, 1980; Hughes, Brooks, Graham, Sheen, and Dickens, 1982; Wiekhorst and Killion, 1986; Robinson, Eubanks, and Eddowes, 1981; McGuinness, Bouwman, and Puig, 1982; Leeds, Raspotnik, and Gular, 1990; Payne, Hirsch, Semple, Farmer, Spring, Sanders, Wimer, Carter, and Hu, 1976; Jenkins, 1982; for reviews, see Bell and Waag, 1998 and Waag, 1991). Compared to predominantly stand-alone systems of the past, DMO not only affords the ability to train team skills, but also to train larger and more complex portions of the mission. Given that these environments afford the ability to train very different and more varied skills, what can be generalized from historical training effectiveness studies is severely limited.

Some multiplayer simulation research suggests DMO enhances individual and team skills for: (1) F-15 pilots (Houck, Thomas, & Bell, 1991), (2) F-16 pilots (Berger & Crane, 1993; Schreiber, Watz, & Bennett, 2003), (3) Tornado pilots and navigators (Huddlestone, Harris, & Tinsworth, 1999), and (4) pilots, forward air controllers, and ground forces executing close air support (Bell, et al., 1996). F-16 pilots who have flown in a distributed environment have rated DMO as a particularly effective training system for missions involving 4-ship air-to-air employment against multiple enemy aircraft (Crane, Schiflett, & Oser, 2000). F-16 pilots also have reported that both individual skills (such as radar mechanization, communication, and building situation awareness) and team skills (such as maintaining mutual support, tactical execution, and flight leadership) are enhanced by DMO training (Crane et al., 2001). However, a comprehensive examination of objective data showing improvement after training does not yet exist.

Current Work

Of the previously cited DMO training effectiveness studies, Schreiber, Watz, & Bennett (2003) provided the only large sample objective analysis. In their preliminary study, the authors examined 19 teams of F-16 pilots (76 total pilots) who flew pre- and post-test point defense air combat scenarios. Over a five-day DMO training week, each four-ship team was exposed

to, on average, 35 DMO air combat scenarios, employing an average team total of 483 shots against an average team total of 293 threat aircraft. Differences in performance on outcome measures from the pre- to post-test were dramatic: The F-16 teams, on average, allowed 63% fewer enemy strikers to target, achieved 24% more enemy fighter kills, and reduced F-16 mortalities by 68%. Those preliminary results suggest that DMO training is highly beneficial for improving air combat competencies, but the study lacked in-depth objective analysis, pilot opinion data, and expert observer rating data. The current work aims to expand upon the results found in Schreiber et. al (2003) by incorporating all these sources of data and a larger sample size.

The current work seeks to fulfill the following objectives:

1. Perform a large-scale, comprehensive within-simulator DMO training effectiveness evaluation by using pilot opinion data, expert observer rating data, objective data, and an assessment of pilots' knowledge structures.
2. Validate the objective measures collected by using converging data from observed objective measures, expected differences as a function of experience, ratings of team and pilot performance made by subject matter experts (SME), and pilot opinion data collected on surveys.

Hypotheses

1. In addition to the subjective literature already discussed, participants and observers of DMT routinely informally report that substantial learning is taking place. We therefore hypothesize that significant improvements in the Monday to Friday benchmark comparison will be observed for a number of indices collected by the Performance Effectiveness Tracking System (PETS) system.

2. We hypothesize that we will not observe a significant trade-off in the observed Monday to Friday performance. That is, pilots will demonstrate improved performance on both offensive *and* defensive skill-related measures.

3. We hypothesize that significant Monday to Friday benchmark improvements will also be observed in the subjective rating data and will corroborate the objective results. However, due to the issues previously discussed, we anticipate the subjective results to be less sensitive for delineating individual constructs.

4. We hypothesize that analysis of the pilot reaction data will also suggest learning benefits as a function of DMO training, thereby providing ecological validity.

5. We hypothesize that there will be a change in pilots' knowledge structures about responsibilities, tasks, and duties pre and post DMO training.

METHODS

Four different kinds of data were collected on the pilots that took part in the training week: PETS, Pathfinder, DMO reactions, and SME evaluations. Each of these are described in detail below.

PETS

The PETS program gathers many variables from the network about the performance of the pilots. However, due to the nature of air combat data, only a limited number of descriptive statistics will be able to be reported in terms of percent change. These variables are described below.

Enemy bombers to reach base
This is the number of enemy bombers that come within 2 nautical miles of the target that the F-16s are defending.

Threat mortality
Out of 8 aircraft (6 fighters and 2 strikers) that are present in each benchmark, the number that are killed by the F-16s.

Viper mortalities
Of four possible, viper mortalities is the number of F-16 mortalities, not including fratricides.

Proportion of viper shots resulting in a kill
Of the missiles shot by the vipers, this is the proportion that resulted in an enemy kill.

Range at launch of radar missile
This is the distance from the F-16 to the enemy at the time that the missile is launched.

Time within critical ranges to threat fighter aircraft

This is the time inside a predetermined range where there is an increased probability of viper mortality.

Proportion of threat missiles resulting in a kill
Of missiles fired by enemy planes, this is the proportion that result in a kill of one of the four F-16s.

Pathfinder
Pathfinder is a numerical and graphical method for describing relations among constructs. Pilots rate by a paired-comparison method the similarity of all pairs of constructs in a set. Using the proximity matrix of similarity ratings, a Pathfinder analysis estimates distances among pairs of constructs. The matrix of an individual or the matrix obtained by averaging ratings from many respondents may be analyzed. Two parameters, q and r, constrain a specific analytic outcome. Parameter q determines the complexity of networks, i.e., the number of paths connecting constructs, and may be set between 2 and n-1, where n is the number of constructs rated. The number of paths estimated increases as the value of q increases. Parameter r determines the exponential power of the algorithm that computes path distances, and for ordinal matrices, always has a constant value of infinity. A measure of coherence indicates the level of internal consistency for a particular matrix of similarities, while a measure of correlation indicates consistency between two different matrices. With respect coherence, values of .20 or less are seen as evidence that participants may not take the task seriously (Pathfinder Getting Started Manual, p. 6).

For the purposes of the present study, individual distance matrices of the pilots were sorted into four groups: (1) Viper 1 pilots beginning of the week (22 matrices), (2) All other Vipers beginning of the week (109 matrices), (3) Viper 1 pilots end of the week (22 matrices), and (4) All other vipers end of the week (109 matrices). For each group the average proximity matrix was computed and analyzed with q set at 2.

DMO reaction
The DMO reaction sheet is filled out by the pilots after the end of training on the last day. It is used to gather feedback about the opinion of the pilots on the utility of the simulator training. The response scale ranged from 1 (Strongly Disagree) to 4 (Strongly Agree). Some questions were worded such that a positive reaction response would require a low score to ensure that the participants were reading the questions, and not simply circling similar numbers for all the questions.

SME ratings
The DMT gradesheet includes 57 broad indicators of 4-ship team performance. The subject matter experts who serve as raters are experienced Air Force pilots. For the time period under study, there were 7 experienced F-16 pilots with a mean of 1621 F-16 hours, and a mean of 19.92 years of service. Raters watched the engagements and rated the overall team performance, in real time, on a relative five point numeric scale. The ratings ranged from 0 – Performance indicates a lack of ability or knowledge, to 4 – performance reflects an unusually high degree of ability. The questions are broken into three groups reflecting the three different areas of the mission. There are 8 questions about the quality of the brief, 40 questions about the engagement, and 9 questions about the quality of the debrief.

Participants
From January, 2002 to May, 2003 there were 35 teams who participated in five-day training research with benchmark exercises at the Mesa DMT site. Additional teams participated in DMO during this time period, but they experienced significantly different syllabi and were therefore not considered potential groups for inclusion in the current study. The mean number of hours in an F-16 was 964 (range 448 to 2088).

Training Research Protocol
During the data collection period, pilots flying the DMO F-16 simulators at the Mesa Research Site flew one of four very similar syllabi, each syllabus consisting of nine "sessions". Each session entailed a one hour brief, an hour of flying, and an hour and a half debrief. There were two sessions each day of the five-day training week, except Friday when the participants had only one session. The syllabi scenarios consisted of 4 v X presentations, both offensive and defensive. Scenarios were designed with trigger events and situations to specifically train certain MEC skills. These syllabi were developed with traditional methods of increasing complexity across a spectrum of probable air-to-air missions and threats.

Each syllabus began with a familiarization session to orient pilots to DMO simulator environment specifics, such as visual ID characteristics and any switchology differences due to F-16 block number or F-16 mission software. Since the simulator layout closely resembled the actual aircraft and since all the declarative and procedural knowledge to be operationally qualified to fly the F-16 had been learned by participants before arriving at HEA, the participants required little familiarity training. Therefore, after the familiarity session, performance increases observed throughout

the course of the subsequent sessions are the result of learning how and when to best employ the skills they had been taught during their Air Force career.

The second session (after the familiarization period) was a benchmark session used to measure pre-DMO training team performance. The Monday benchmark session consisted of flying point defense engagements with three "benchmark scenarios." All benchmark point defense scenarios were 4 v 8 (6 hostiles and two strikers) and were designed to be equally complex according to the absolute complexity scoring scheme outlined by Denning, Bennett, and Crane (2002). Five point defense benchmark scenarios were developed and the complexity analysis revealed that they all were indeed equally complex. For the Friday benchmark session, participants flew, in the same flight/cockpit assignment, the mirror-image of the three benchmarks that were flown on Monday.

There were five point defense benchmark scenarios, each with a mirror-image scenario (ten total) available. The goal of the benchmark mission is to prevent the bombers from reaching the base – success being striker denial or kill. All benchmark scenarios have been judged to be equally difficult (Denning, Bennett, & Crane, 2002). These scenarios were selected for examination in the present study because: (1) all the benchmark engagements have equivalent levels of complexity, (2) three benchmark scenarios occur at the beginning and end of the week-long DMT syllabus, (3) the same pilots perform the benchmark scenarios in the same team positions at the beginning and end of week, and (4) the benchmarks were flown under real-time kill removal and strict data collection rules.

The building-block training was initiated during session three and continued through the course of the week. During sessions three through eight, participating teams were exposed to four to eight full engagements, which began at the CAP or push point and generally concluded with a logical conclusion such as "Bingo" fuel, "Winchester" ordnance, all threats killed or multiple friendly losses. While these training sessions emphasized Defensive Counter Air scenarios (DCA), pilots also flew Offensive Counter Air (OCA) and air-to-ground missions. All engagements were flown versus simulation of actual threat aircraft, air-to-air ordnance and surface-to-air ordnance. These 30+ training engagements provided a very rich environment for air-to-air training and were the equivalent of flying more than ten friendly four-ship missions, with each mission opposed by 8 - 16 dissimilar adversary aircraft, or over 400 total sorties. The training sessions also provided real-time enemy kills (complete with visual explosions) and real-time friendly losses. Successive training sessions progressed in difficulty with increases in number of threat aircraft, type of threat aircraft, threat aircraft reactivity/maneuver or an increase in vulnerability time.

RESULTS

PETS

31 teams had valid PETS data available for analysis. Due to the nature of air combat data, only a limited number of descriptive statistics will be able to be reported in terms of percent change. All statistics showed improvements in the expected direction. Compared to the Monday benchmarks, the Friday benchmarks showed 69% fewer viper mortalities, 61% fewer enemy bombers reaching base, 25% more threats killed, 10% longer range at launch of missile, 69% less time spent in critical range by F-16s, 55% fewer threat shots resulting in a kill, and 7% more viper shots resulting in a kill.

Pathfinder

Thirty-five teams completed the pathfinder program before and after a training week. An examination of the estimates of coherence for the four conditions in Table 1 shows that the aggregate levels of coherence exceed a value of .20, a finding that suggests average matrices capture systematic information about distances among the 15 constructs in the set rated. Note also that there is a distinct increase in coherence from the beginning of the week to the end of the week. Because these pilots maintained the same Viper positions at the beginning and end of the week, the increase in coherence is most likely the result of training and other activities that took place during the DMO experience.

Table 1. Estimates of Coherence for Average Proximity Matrices

Pilots	Time of Week	
	Beginning	End
Viper 1	.486	.568
All Other Vipers	.481	.535

An examination of the estimates of correlation between matrices for these four conditions in Table 2 shows consistently very high levels of association, indicating very few differences in the rank order of average values of proximity matrices among the four groups.

Table 2. Estimates of Correlation[1] Among Average Proximity Matrices

Classification of Matrix	All Other Vipers Time 1	All Other Vipers Time 2	Viper 1 Pilots Time 1
All Other Vipers Time 2	.986		
Viper 1 Pilots Time 1	.964	.955	
Viper 1 Pilots Time 2	.964	.963	.957

DMO reaction

175 pilots completed the DMO reaction form. After reversing the scores of the questions that are worded negatively, the overall grand mean was 3.61 (on a scale of 1 to 4), with a standard deviation of 0.323, for the 57 questions analyzed.

SME ratings

The gradesheet data were grouped according to area of the mission resulting in three variables for analysis: brief, engagement, and debrief. The ratings of the benchmark scenarios on Monday (pre-training) and Friday (post-training) were compared using three repeated measures t-tests. The results showed that the mean rating for the quality of the brief increased from Monday (M=1.52, SD=0.73) to Friday (M=2.65, SD= 0/49), t=7.546, p<0.001, indicating a higher quality brief after training. The mean rating of the engagement also increased from Monday (M=1.27, SD= 0.34) to Friday (M=2.39, SD= 0.46), t=11, p<0.001, indicating better performance after a week of simulator training. In addition, the mean rating of the quality of the debrief increase from Monday (M=1.59, SD= 0.74) to Friday (M=2.71, SD = 0.51), t=7.25, p< 0.01, indicating a higher quality debrief after training.

DISCUSSION

The results just outlined provide the first set of converging results from multiple sources showing a measurable benefit from simulator training. Previous studies of training effectiveness have focused on one or two subjective methods of assessing learning. In the current study training research effectiveness was measured with four diverse methods, both objective and subjective. The ratings given by the SMEs of the team performance increased from Monday to Friday. The Pathfinder data showed more coherence on Friday than on Monday. The DMO reaction forms showed a positive experience from the pilot's perspective. And, the objective PETS data show changes consistent with better performance on Friday than on Monday. Therefore, all four methods showed changes in a direction reflective of positive training during the training week. Taken together, these results indicate that the pilots are receiving significant benefit from their time spent in the simulators.

In the introduction, we listed four hypotheses about the results of the study. The results are listed below:

1. The PETS data showed significant improvement from Monday to Friday.

2. The PETS data showed improvements in both offensive and defensive related skill measures after a week of training.

3. The subjective SME rating data showed higher ratings on Friday than on Monday.

4. The pilots expressed, on the DMO reaction survey, a positive learning experience during the week.

5. Pilots' knowledge structures about responsibilities, tasks, and duties changed, and became more coherent after DMO training.

These data represent an extension of previous preliminary data from this lab that were presented in Watz, Schreiber, Keck, McCall, & Bennett, 2003. That data presented preliminary results from the PETS system. This paper has expanded on those results with more PETS data, and included other data sources to support the hypotheses of training effectiveness. In addition, a more extensive report is in preparation that will include more data, especially objective process and skill metrics.

Of the constructs that were analyzed in this study, some were measures of individual performance and some were measures of performance of the overall four ship team. Although this is two different ways of

measuring performance, both methods lend support to training effectiveness.

The conclusions from this study lend overall support to simulator training effectiveness. No measure of performance effectiveness is perfect, and although each individual construct used in this study may have its own individual flaws, when viewed as a whole, there is great support for training effectiveness.

Much work remains to be done to definitely establish the effectiveness of simulator training of combat pilots. Future work in this area will involve expanding the methods used to assess pilot performance. New measures and further analysis of current measures will provide more support for our conclusions. In addition, knowledge about the length of time that the pilots retain the skills trained in the simulator will be examined by measuring the skills at different points in time following training.

ACKNOWLEDGEMENTS

The views expressed in this paper are those of the authors and do not necessarily reflect the views or the U.S. Air Force, Lockheed Martin, or Boeing. The authors would also like to thank, in alphabetical order, Jim Marx, Toni Portrey, Bart Raspotnik, Robbie Robbins, Margaret Sallay, Traci Smith, Don Smoot, Dr. Bill Stock, and Major Steve Symons.

REFERENCES

Bell, H. H., Dwyer, D. J., Love, J. F., Meliza, L. L., Mirabella, A., & Moses, F. L. (1996). *Recommendations for planning and conducting multi service tactical training with distributed interactive simulation technology (A Four-Service Technical Report).* Alexandria, VA: US Army Research Institute.

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.

Bennett, W., Jr., & Crane, P. (2002). *The deliberate application of principles of learning and training strategies within DMT.* Paper presented at the NATO SAS-038 Working Group Meeting, Brussels, Belgium.

Berger, S., & Crane, P. M. (1993). Multiplayer simulator based training for air combat. In *Proceedings of 15th Industry/Interservice Training Systems Conference*, Orlando, FL: National Security Industrial Association.

Carolan, T., MacMillan, J., & Schreiber, B.T., (2003). Integrated Performance Measurement and Assessment in Distributed Mission Operations Environments: Relating Measures to Competencies. Paper presented at the *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2003)*

Crane, P. M., Schiflett, S. G., & Oser, R. L. (2000). Roadrunner '98: *Training effectiveness in a Distributed Mission Training exercise* (AFRL-HE-AZ-TR-2000-0026, ADA391423). Project 2743. NTIS (2000).

Crane, P., Robbins, R., Bennett, W., Jr., & Bell., H. H. (2001). Mission complexity scoring for distributed mission training. Paper presented at *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC 2001), Orlando, FL.*

Colegrove, C. M., & Alliger, G. M. (2002). Mission essential competencies: Defining combat 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.

Denning, T., Bennett, W., Jr., & Crane, P. (2002). Mission complexity scoring in distributed mission training. Paper presented at the *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2002).*

Gray, T. H., Chun, E. K., Warner, H. D., & Eubanks, J. L. (1981). *Advanced flight simulator: Utilization in A-10 conversion and air-to-surface attack training* (Rep. no. AFHRL-TR-80-20, AD A094608). Williams AFB, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Gray, T. H., & Fuller, R. R. (1977). *Effects of simulator training and platform motion on air-to-surface weapons delivery training* (Rep. No. AFHRL-TR-77-29, AD A043 648). Williams AFB, AZ: Air Force Human Resources Laboratory, Operations Training Research Division.

Hagin, W. V., Dural, E., & Prophet, W. W. (1979). *Transfer of training effectiveness evaluation: US Navy device 2B35* (Seville Research Corporation

Rep. No. TR 79-06). Pensacola, FL: Chief of Naval Education and Training.

Houck, M. R., Thomas, G. S., & Bell, H. H. (1991). *Training evaluation of the F-15 advanced air combat simulation* (Rep. No. AL-TP-1991-0047, AD A241 674). Williams AFB, AZ: Armstrong Laboratory, Aircrew Training Research Division.

Huddlestone, J., Harris, D. & Tinsworth, M. (1999). Air Combat Training - The Effectiveness of Multi-Player Simulation. Paper presented at *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC 1999)*, Orlando, Florida.

Hughes, R., Brooks, R. B., Graham, D., Sheen, R., & Dickens, T. (1982). Tactical ground attack: On the transfer of training from flight simulator to operational Red Flag exercise. In *Proceedings of the 4th Interservice/Industry Training Equipment Conference: Volume I* (pp. 127-130). Washington DC: National Security Industry Association.

Jenkins, D. H., (1982). Simulator training effectiveness evaluation. Tactical Fighter Weapons Center Nellis AFB NV (ADB068021)

Kellogg, R., Prather, E., & Castore, C. (1980). Simulated A-10 combat environment. In *Proceedings of the Human Factors Society 24th Annual Meeting* (pp. 573-577). Santa Monica, CA: Human Factors Society.

Krusmark, M. A, Schreiber, B. T. & Bennett, W., Jr. (submitted). *The effectiveness of a traditional gradesheet for measuring air combat team performance in simulated distributed mission operations.* Submitted for Technical Report.

Leeds, J., Raspotnik, W. B., & Gular, S. (1990). *The training effectiveness of the simulator for air-to-air combat* (Contract No. F33615-86-C-0012). San Diego, CA: Logicon.

Lintern, G., Sheppard, D. J., Parker, D. L., Yates, K. E., & Nolan, M. D., (1989). Simulator design and instructional features for air-to-ground attack: A transfer study. *Human Factors, 31*, 87-99.

McGuinness, J., Bouwman, J. H., & Puig, J. A. (1982). Effectiveness evaluation for air combat training. In *Proceedings of the 4th Interservice/Industry Training Equipment Conference: Volume I (pp. 391-396)*. Washington, DC: National Security Industrial Association.

Payne, T. A., Hirsch, D. L., Semple, C. A., Farmer, J. R., Spring, W. G., Sanders, M. S., et al. (1976). *Experiments to evaluate advanced flight simulation in air combat pilot training: Vol1. Transfer of learning experiment.* Hawthorne, CA: Northrop Corporation.

Robinson, J. C., Eubanks, J. L., & Eddowes, E. E., (1981). *Evaluation of pilot air combat maneuvering performance changes during TAC ACES training.* Nellis AFB, NV: U.S. Air Force Tactical Fighter Weapons Center.

Schreiber, B.T., Watz, E.A, & Bennett, W., Jr. (2003). Objective Human Performance Measurement in a Distributed Environment: Tomorrow's Needs. Paper presented at the *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2003)*.

Schreiber, B.T., Watz, E.A, 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.

Schvaneveldt, R. W. (1998). Info.doc. [Computer program documentation file]. Gilbert, AZ: Interlink.

Symons, S, France, M., Bell, J., & Bennett, W., Jr. (2003). *Linking Knowledge and Skills to Mission Essential Competency-Based Syllabus Development for Distributed Mission Operations*, Unpublished paper from Air Force Research Laboratory, Human Effectiveness Division, Warfighter Training Directorate, Mesa, AZ.

Waag, W. L. (1991). The value of air combat simulation: strong opinions but little evidence. In, *Proceedings of the Royal Aeronautical Society Conference on Flight Simulation and Training.* pp London: Royal Aeronautical Society.

Watz, E. A., Schreiber, B. T., Keck, L., McCall, M., & Bennett, W., Jr. (2003). Performance measurement challenges in distributed mission operations environments. Paper presented at *the Simulation Interoperability Workshop Conference*, Orland, FL.

Wiekhorst, L. A., (1987). *Contract ground-based training evaluation.* Executive summary. Langley AFB, VA: Tactical Air Command.

Wiekhorst, L. A., & Killion, T. H. (1986). *Transfer of electronic combat skills from a flight simulator to the aircraft* (Rep. No. AFHRL-TR-86-45, AD C040 549). Williams AFB, AZ: Air Force Human Resources Laboratory, Operations Training Division.