Assessing High-Fidelity Training Capabilities Using Subjective and Objective Tools

Leah J. Rowe
L-3 Communications
Mesa, AZ
leah.rowe@mesa.afmc.af.mil

Justin H. Prost
Lumir Research Institute
Mesa, AZ
justin.prost@lumir.com

Brian T. Schreiber
Lumir Research Institute
Mesa, AZ
brian.schreiber@lumir.com

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

ABSTRACT

Instructors often assess training effectiveness using subjective evaluation tools. The use of evaluation by Subject Matter Experts (SMEs) assumes that the experts can distinguish between small but meaningful differences in the measured domain. Subjective evaluations by experts provide both an efficient and effective means of identifying the strengths and weaknesses of the assessed entity. In the area of simulation development, SME assessments evaluate the training capabilities of systems, identify deficiencies, and compare the relative impact of the various deficiencies. This paper presents methods that utilize subjective assessments from SMEs and compares SME ratings of Mission Essential Competency (MEC) experiences with objective performance measures. The methodology entails mapping the correspondence between MECs and objective performance measures. Additionally, we mapped performance measures to training scenarios in order to determine the appropriate skills for evaluation. This study uses performance measures based on the capabilities of the simulators in our laboratory. The congruence of the subjective evaluations by experts and objective simulator performance variables provides validation for the use of subjective assessments completed by experts. The results provide a strong framework for building an understanding of the relationship between subjective and objective performance data to measure training effectiveness.

ABOUT THE AUTHORS

Leah J. Rowe is a Research Scientist with L-3 Communications at the Air Force Research Laboratory, 711th Human Performance Wing in Mesa, AZ. She completed her M.S. in Applied Psychology at Arizona State University in 2007. Leah is presently pursuing a Ph.D. in Industrial/Organizational Psychology at Capella University.

Justin H. Prost is a Research Scientist with Lumir Research Institute. He completed his Ph.D. in Developmental Psychology at Arizona State University in 2001. Recently, he has worked on current simulation research at the Air Force Research Laboratory.

Brian T. Schreiber is CEO and Senior Scientist with Lumir Research Institute in support of the Air Force Research Laboratory, 711th Human Performance Wing, 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, 711th Human Performance Wing, in, Mesa AZ. He received his Ph.D. in Industrial/Organizational Psychology from Texas A&M University in 1995.

Assessing High-Fidelity Training Capabilities Using Subjective and Objective Tools

Leah J. Rowe
L-3 Communications
Mesa, AZ
leah.rowe@mesa.afmc.af.mil

Justin H. Prost
Lumir Research Institute
Mesa, AZ
justin.prost@lumirresearch.com

Brian T. Schreiber
Lumir Research Institute
Mesa, AZ
brian.schreiber@lumirresearch.com

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

INTRODUCTION

Assessment systems, training programs, and subjective assessment tools are the product of expertise. To become an expert, one must obtain both skills and knowledge in a specific domain (Schvaneveldt, Tucker, Castillo, & Bennett 2001). We rely on subject matter experts (SMEs) in many fields (e.g., law enforcement, human factors, medicine, and engineering). The military is no exception to this rule, and uses SMEs regularly.

SMEs have knowledge, skills, and experiences that set them apart from the average field practitioner. They can identify subtle cues that less-experienced operators may miss during complex tasks and in specific environments. SMEs often provide simple assessment solutions for very complex measurement tasks (Schreiber, Gehr, & Bennett, 2006).

Yet even a SME, may find it difficult to assess performance effectively. Historically, Warfighter performance has been assessed using subjective grading measures either by SMEs or Instructor Pilots (Schreiber, et al., 2006; Krusmark, Schreiber, & Bennett, 2004; Crane, Robbins, & Bennett, 2000). Researchers continually strive to identify or create objective performance measures. At the Air Force Research Laboratory (AFRL) in Mesa, Arizona, researchers have developed a system that collects objective data from a complex high-fidelity simulation environment. This paper discusses a method of combining objective and subjective data to assess training research in the Distributed Mission Operations (DMO) Training Research Testbed (TRT) at AFRL Mesa.

We begin by discussing the differences between subjective and objective data, and highlight the advantages of each. Next, we discuss the AFRL DMO TRT highlighting the approach that combines subjective and objective data to create a metric to measure training effectiveness. Finally, we discuss the methodology used, findings, and implications for the future.

Subjective versus Objective Performance Assessment

Subjective Data

Subjective data provides the only means for assessing both opinions and preferences. Subjective data is collected frequently as it is typically easy to obtain and inexpensive, these two factors may influence practitioners when they select a data collection method (Cushman & Rosenbery, 1991). Nevertheless, in some situations subjective data is the only data source that is available or feasible.

At the DMO TRT, we collect both subjective and objective performance data. F-16 SMEs generate the subjective data by completing SPOTLITE (Scenario-based Performance Observation Tool for Learning in Team Environments). SPOTLITE allows observers to measure and assess team and individual performance in live and simulated training exercises in real time (MacMillan, Entin, Morely, & Bennett, under review).

Objective Data

Researchers often prefer objective data in research, because it ideally lacks bias; however, it is often difficult to obtain. To be truly objective, there must be an "absolute" answer absent of human opinion. This situation in itself creates a barrier when building objective assessments. In addition, objective measures are generally more costly and time consuming than subjective measures (Cushman & Rosenbery, 1991).

In the DMO TRT, we collect objective performance data with the Performance Evaluation Tracking System (PETS). PETS provides the Warfighter with exact data regarding their actions during live and training events

by collecting and distilling millions of data points directly from the simulator (Schreiber & Bennett, 2006). We describe PETS in more detail below.

Which Assessment Method to Use?

PETS gathers micro-data that is not feasible for a human to track, whereas SPOTLITE assesses performance with criteria that only a SME can assess. It is necessary to identify the most appropriate assessment method for any performance evaluation. The fundamental differences between PETS and Spotlite make it clear that performance assessment does not fall in a "one size fits all" category.

Subjective assessments often prove to be the most efficient mechanism for obtaining information; however, when subjective assessments are appropriate, it is important to assure data quality by gathering it from a reliable source. SMEs have expertise that improves the reliability of subjective data

In prior research, objective data showed that, F-16 pilot performance improved from pre- to post-training in the DMO TRT (Schreiber & Bennett, 2006; Rowe, Gehr, Cooke, & Bennett, 2007). Additionally, subjective measures showed that pilot knowledge changed from pre- to post-training in the DMO TRT as well (Rowe, Gehr, Cooke, & Bennett, 2007; Rowe, Schvaneveldt, & Bennett, 2007).

This paper presents an approach to mapping subjective F-16 SME ratings to objective performance data. Building a process that integrates SME evaluations and objective performance data will allow integration of more sophisticated training protocols in the DMO environment. In any training environment, SMEs are limited to what they can observe. The DMO TRT has more performance information available, a result of both technological advances (e.g. objective performance measurement tools) and the increased number of participants. Providing instructors with objective performance measures will allow development of more effective and efficient training protocols. One such example is the development of "adaptive training."

Distributed Mission Operations Training Research Testbed

DMO Defined

DMO is a system of networked simulators that supports multi-player training for combat exercises. DMO is different from stand-alone simulation systems, such as those used to train emergency procedures, in that it provides combat-like experiences involving real-time interaction with other entities, both virtual (e.g., a flight wingman in another simulator) and constructive (e.g., hostile entities). The objective of DMO is to train higher-order skills and improve team coordination while executing significant portions of an entire mission (Colegrove & Alliger, 2002).

The DMO TRT consists of four high-fidelity F-16 simulators, a high fidelity Air Battle Manager Simulator, a computer-generated threat system, and an instructor/operator station. The DMO TRT also includes a well equipped brief/debrief room (the DMO TRT is shown in Figure 1).

Figure 1. Overall view of Mesa AFRL DMO Training Research Testbed

Mission Essential Competencies

Syllabi trained in the DMO TRT are structured based on Mission Essential Competencies (MECs), defined as "higher-order individual, team, and inter-team competencies that a fully prepared pilot, crew or flight requires for successful mission completion under adverse conditions and in a non-permissive environment" (Colegrove & Alliger, 2002, p. 12). A competency-based training structure defines a standard level of proficiency or competency that one must have in order to be efficient in his/her job, thus emphasizing ways to address deficiencies in skills, knowledge, or experience in individuals, teams, or crews (Schreiber & Bennett, 2006).

Performance Evaluation Tracking System

PETS developed at AFRL, as an Advanced Technology Demonstration for the Air Combat Command, is a software tool that enables multi-platform, multi-level measurement at the individual, team, and inter-team levels in complex, live, virtual, and constructive environments (Schreiber & Bennett, 2006).

Installed in the DMO TRT PETS collects, stores, and organizes up to one million data points per minute. Schreiber and Bennett (2006) validated the use of PETS in a simulated environment. Additionally, they were able to define the most sensitive air-to-air measures for the F-16 in this environment, meaning the measures that are most significantly impacted from pre- to post training in the DMO TRT.

METHODS

Participants

Two hundred-seventy-two F-16 fully qualified F-16 pilots from United States Air Force, Air National Guard, and Air Force Reserve pilots participated in this study. The pilots consisted of 53 teams or four or five pilots each. Their mean age was 33.1, and they had an average of 10.8 years of military service and 1,016 F-16 flight hours.

Another sample consisted of seven F-16 SMEs. All participants were male, with a mean age of 40.8 years. Two are active in the Air National Guard and five retired from the Air Force between one and two years ago.

Procedures

DMO Training Research Week
Each team participated in nine 3½-hour training sessions over the course of the single DMO training week. Each session included a one-hour briefing, an hour of flying multiple engagements of the same mission genre, and a 90-minute post-mission debrief. Syllabus scenarios were either offensive or defensive, and consisted of four F-16s versus a varying number of threats. The team flew three benchmark scenarios at the beginning of the week and again at the end of the week for evaluation purposes.

Flight Performance
We assessed flight performance using PETS. Metrics were derived to measure performance change in three areas: weapons employment, weapons engagement zone management, and overall performance.

The benchmarks were constructed as scenarios where the four-ship of F-16s and their Air Battle Manager defended against eight threats (six hostiles and two strikers). All benchmarks were designed to be of equal complexity. We randomly assigned each team three-benchmark scenarios. The participants flew in the same cockpits during all benchmark scenarios. On day five, teams flew mirror image missions of the three benchmarks. Figure 2 illustrates a benchmark and its mirror image. All of the benchmark scenarios that were utilized during this research are equally complex (Denning, Bennett, & Crane, 2002).

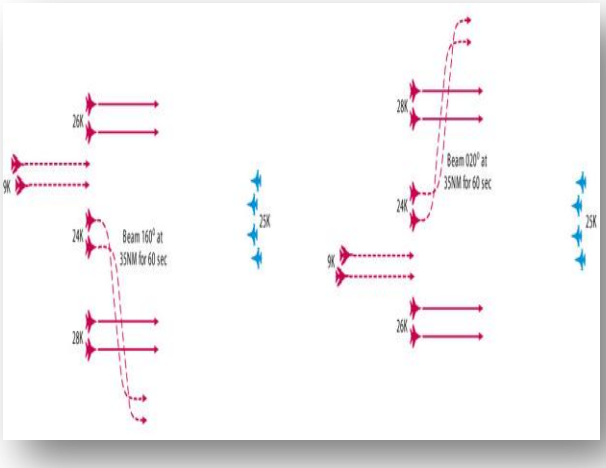

Figure 2. Mirror-Image Point Defense Benchmark Scenarios

Knowledge, Skill, and Performance Mappings
F-16 SMEs completed three sets of ratings to complete the tasks described in the following paragraphs. Each task utilized an identical Likert scale (0 = Not Relevant, 1 = Somewhat Relevant, 2 = Largely Relevant, and 3 = Extremely Relevant).

For the first measure, seven SMEs each completed 36 rankings mapping the relevance of all knowledge areas and skills defined in the air-to-air MECs (Colegrove & Alliger, 2002) to our benchmark scenarios.

For the second measure, four SMEs each completed 1,739 ratings of the relevance of all conceptual performance measures to the air-to-air knowledge areas and skills defined in the air-to-air MECs (Colegrove & Alliger, 2002).

The final set of ratings mapped the relevance of objective conceptual performance measures (developed as part of a Performance Measurement Workshop) to objective PETS measures. For this task, seven F-16 SMEs each completed 2,194 ratings.

ANALYSES

We designed the analyses to identify the correspondence between objective performance measure and subjective evaluations provided by SMEs.

Step One: In step one we calculated the average for the ratings for MEC knowledge areas and skill relevance to benchmark scenarios (measure 1) across the SMEs. These ratings provided the basis for organizing those skills and areas of knowledge based on relevance to the benchmark scenarios.

Step Two: In this step, we combined the ratings identifying the degree to which the MEC knowledge and skills are involved in the benchmark scenarios with the ratings evaluating the relationship between the MEC knowledge and skills and the conceptual performance measures. The new scores represent the relationship of the MEC knowledge and skills to the conceptual performance scores, weighted by the degree to which the benchmark scenarios capture each of the MEC knowledge and skill areas. The sum for each PETS conceptual measure is computed to represent the degree to which each conceptual measure is influenced by the MEC knowledge and skills trained on the benchmark scenarios.

Step Three: Based on the SME subjective assessments step three determined the degree to which each metric influences benchmark scenarios. We multiplied the scores derived in step two by the ratings from the mapping between the conceptual measures and the metrics (step one). The resulting values represent the relationship between the conceptual measures and the metrics, weighted by the degree to which those measures would be trained on benchmark scenarios. Finally, these values were summed across the conceptual measures for each metric, resulting in a single value for each metric.

Step Four: Step four identified the PETS performance measures that improved across DMO training research week. We entered the metrics in the three areas of interest into the data set with the value that represented the proportion of improvement on the metric over the week. Improvement is defined as an increase or decrease in the metric, depending on the desired outcome (e.g. "shortest distance of a striker to base" showed improvement by a percent increase in that distance).

Step Five: In step five, we computed Pearson product-moment correlation coefficients between the objective performance measures from training weeks and the scores for MEC knowledge areas and skills involved in benchmark training, according to subjective evaluations.

RESULTS

For the analysis of the ratings relating MEC knowledge areas and skills to the benchmark scenarios (computed in step 1) the average knowledge rating for the benchmark scenarios was 2.45, with a standard deviation of 0.50. The average skill rating for the benchmark scenarios was 2.66, with a standard deviation of 0.30. The SMEs rated both the MEC knowledge area and skills with average ratings between approximately 1.5 and the maximum of 3. This range in scores indicates the high level of relevance of the benchmarks to the knowledge and skills necessary for pilot readiness, while still being able to discriminate between more and less relevant skills and areas of knowledge; Table 1 presents the top five MEC knowledge areas and skills.

Table 1. Top five MEC Knowledge Areas and Skills

Top 5 MEC Knowledge Areas
1. Mission Objectives
2. Threat Capabilities
3. Communication Standards
4. Commit Criteria
5. Formation

Top 5 MEC Skills
1. Builds Picture
2. Listens
3. Multitasks
4. Radar Mechanization
5. Sorts Targets

The second step generated scores that provided an indication of the relevance of each PETS conceptual measure to the benchmark scenarios. We computed an average score for knowledge areas and skills for each conceptual performance measure. There are 12 MEC knowledge areas and 24 MEC skill areas. The average score for MEC knowledge across the conceptual performance measures is 1.89, with a standard deviation of 0.88. The average score for MEC skills across the conceptual performance measures is 2.42, with a standard deviation of 1.07. There are 44 conceptual performance measures in this study. Table 2 illustrates the top five conceptual performance measures influenced by MEC knowledge areas and skill for the benchmark scenarios.

Table 2. Top five Conceptual Performance Measures for MEC Knowledge Areas and Skills

Top 5 Conceptual Performance Measures for MEC knowledge
1. How close red came to point/area/HVAA
2. Number of visual merges with second red within factor range
3. Fly into frag
4. Air-to-air shot measures
5. How many times painted by red air radar

Top 5 Conceptual Performance Measures for MEC skills
1. Quality of communications
2. Mutual support
3. Number visual merges with second red within factor range
4. Percent of red air targeted by targeting range
5. Percent of red air detected by min targeting range

During the third step, we calculated a weighted score representing the degree to which each of the PETS performance measures should improve based on the SME subjective assessments. To identify the degree to which each of the PETS metrics included in the current study would change based on subjective assessments, the relevance of each of the metrics to training benchmark scenarios. The average knowledge score across PETS metrics for this step was 2.09, with a standard deviation of 0.38. The average skill score across PETS metrics for this step was 3.00, with a standard deviation of 0.48.

In the fourth step, we identified seventeen performance measures from PETS to include in the current analyses. We extracted the percent improvement for each metric, based on change over the week to the end of the training week. Table 3 shows the top five and bottom five rank ordered measures.

Table 3. Top five and bottom five metrics showing improvement

Top 5 Metrics
1. Bombers killed before reaching base
2. Average N-Pole Exposure Time
3. Bombers reaching base
4. MAR-1 time for team
5. MAR time for team

Bottom 5 Metrics
5. MOR time for team
4. Slant range to target (AAMRAM) at launch
3. 2D range to target (AAMRAM) at launch
2. Proportion of all threats killed
1. Proportion of Viper shots resulting in kill

The final step compared the degree to which pilots improved on different objective performance measures with the anticipated improvement on the measures, based on the subjective SME assessments. A correlation between the scores from MEC knowledge areas and the percent improvement was not significant, $r(15) = 0.23$, n.s. The correlation between the scores from MEC skills and the percent improvement was not significant, $r(15) = 0.20$, n.s. In order for a correlation to be significant with 15 degrees of freedom the value of the coefficient would need to be .48.

DISCUSSION

Our findings provide preliminary support for further development of the process presented here. Identifying the areas in which subjective and objective performance measurements are most effective and efficient offers a powerful tool for developing and refining training programs. Additionally, the correspondence between subjective and objective performance measures that we report here would enable instructors to select and integrate objective performance measures into training. For example, if an instructor sees that a pilot is not improving on certain objective performance metric, they can use the correspondence to know which MEC skills and knowledge should areas should be remediated in training. Additional investigations will refine the process to provide a more rigorous closed-loop, adaptive training process.

The lack of significant correlations between the subjective scores and the objective improvements should not be interpreted as a lack of evidence for the process. Although the correlations were not found to be significant, only 17 PETS metrics were used in the current study, providing few degrees of freedom. The correlation coefficients, though in the range of small relationships, were both in the correct direction and represent small effect sizes.

In addition to the small number of metrics included in this study, this is the first time that this rating system for mapping measurement frameworks has been used in this environment and is still in the testing phase of the development process. The knowledge, skill, and performance mappings were done with a small sample size to provide enough data to validate the process. An

increase in the number of SMEs providing ratings for mappings may provide for sensitive measures, decreasing the variability and improving the relationship between the objective and subjective performance measures.

Although the findings could have been stronger for validating the relationship between objective and subjective performance measures, the results of the process do provide a strong framework for building an understanding of the relationships. The use of objective performance data in the training environment will ultimately be limited on the ability of instructors and trainees to disseminate and understand the feedback from the objective measurement systems.

The process presented in the current framework can be used to develop more sophisticated competency-based training environments. Furthermore, once the process explored in this study is validated the metric can be used as an assessment tool in an adaptive training environment. Future research might investigate the full range of available objective performance metrics and the impact of system fidelity on the mapping process. Finally, the next goal of the current research will be to integrate this work as an additional tool for enhancing training environments.

ACKNOWLEDGEMENTS

This research was performed at the Air Force Research Laboratory, Warfighter Readiness Research Division in Mesa, AZ, under Air Force contract 8650-05-D-6502, Principle Investigator Dr. Winston Bennett, Jr.

REFERENCES

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

Crane, P. M., Robbins, R., & Bennett, W. J. (2000). Using Distributed Mission Training to Augment Flight Lead Upgrade Training. *2000 Interservice/Industry Training, Simulation and Education Cnference (I/ITSEC) Proceedings.* Orlando, FL: National Security Industrial Association (AFRL-HE-AZ-TR-2000-0111, ADA394919). Proj 2743. F41624-97-D-5000. Mesa, AZ: L3 Communications.

Cushman, W. H., & Rosenbery, D. J. (1991). *Human factors in product design.* Amsterdam: Elsevier.

Denning, T., Bennett, W. Jr., & Crane, P. M. (2002). Mission Complexity Scoring in Distributed Mission Training. In 2002 *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. Orlando, FL: National Security Industrial Association.

Krusmark, M., Schreiber, B. T., & Bennett, W. J. (2004). *The Effectiveness of a Traditional Gradesheet for Measuring Air Combat Team Perfomrance in Simulated Distribute Mission Operations.* (AFRL-HE-AZ-TR-2004-090). Air Force Research laboratory, AZ: Warfighter Readiness Research Division.

MacMillan, J., Entin, E. B., Morley, R., & Bennett, W. Jr. *Measuring team performance in complex and dynamic military environments: The SPOTLITE method*. Manuscript in preparation.

Rowe, L. J., Gehr, S. E., Cooke, N. J., & Bennett, W. J. (2007). Assessing Distributed Mission Operations Using the Air Superiority Knowledge Assessment System. *Hman Factors and Ergonomics Annual Meeting.* Baltimore, MD.

Rowe, L. J., Schvaneveldt, R. W., & Bennett, W. J. (2007). Measuring Pilot Knowledge in Training: The Pathfinder Network Scaling Technique. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).* Orlando, FL: National Security Industrial Association.

Schreiber, B. T., & Bennett, W. J. (2006). Distributed Mission Operations Within-Simulator Effectiveness Baseline Study: Summary Report. *(AFRL-HE-AZ-TR-2006-0015-Vol I). Air Force Research Laboratory, AZ: Warfighter Readiness Research Division* .

Schreiber, B. T., & Bennett, W. J. (2006). *Distributed Mission Operations Within-SImulator Training Effectiveness Baseline Study.* Mesa AZ: Air Force Research Laboratory, Warfighter Training Resaerch Division.

Schreiber, B. T., Gehr, S. E., & Bennett, W. J. (2006). Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study: Real-Time and Blind Expert Subjective Assessments of Learning. *AFRL-HE-AZ-TR-2006-0015-Vol II* . Air Force Research Laboratory, AZ: Warfighter Readiness Research Division.