Objective Human Performance Measurement in a Distributed Environment: Tomorrow's Needs

Brian T. Schreiber, Eric A. Watz
Lockheed Martin
Mesa, AZ
Brian.Schreiber@mesa.afmc.af.mil,
Eric.Watz@mesa.afmc.af.mil

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

ABSTRACT

Networked simulation development continues its rapid progress, but automated human performance assessment development has been almost completely neglected. Typically, subject matter expert opinions and surveys are used to assess human performance. While subjective ratings for complex simulation environments provide a valuable assessment for overall performance, a comprehensive effectiveness evaluation should include objective measures for both in-simulator and transfer to actual environment assessments. Furthermore, the individual assessment of numerous skills can exceed the attentional resources a subject matter expert has to offer. Thus, an automated objective skill measurement system is required to properly evaluate the training effectiveness of networked simulations. For example, if one wanted to track the amount of time an aircraft has spent within different "range rings" to a threat, an automated tool can quickly and precisely track this information by "listening" to network traffic and calculating distances from the positional information provided by each entity. An automated performance assessment tool could go beyond simple measures such as kill ratios and weapon hit ratios; tracking hundreds of variables, thereby providing objective assessments for individual and team skills quickly and accurately. This paper presents a general methodology for capturing automated objective assessments from a networked simulation environment. Research in this area reveals that, although many objective assessments can be made today, additions to the current DIS/HLA protocol standards would increase the number of opportunities and methodologies used to measure individual and team performance. This paper presents results showing that human performance assessments for networked simulation is possible and advocates for new requirements to enable a standardized, extensive suite of measures. These measures will allow researchers to quantify the amount of learning that has taken place in a networked simulation environment, in terms of both outcome and process measures, and the standardization enables cross-comparison of effectiveness results.

ABOUT THE AUTHORS

Brian T. Schreiber is a Staff Scientist with Lockheed Martin at the Air Force Research Laboratory, Warfighter Training Research Division, in Mesa, AZ. He completed his M.S. in Human Factors Engineering at the University of Illinois at Champaign-Urbana in 1995.

Eric A. Watz is a Software Engineer with Lockheed Martin at the Air Force Research Laboratory, Warfighter Training Research Division, in Mesa, AZ. He is currently pursuing an M.S. in Computer Information Systems. He completed a B.S. in Computer Science from Arizona State University in 2002 and a B.S. in Human Biology from the University of Wisconsin in 1997.

Winston Bennett, Jr. is a senior research Psychologist and senior scientist for training systems technology and performance assessment at the Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Training Research Division in Mesa, AZ. He received his PhD in Industrial/Organizational Psychology from Texas A&M University in 1995. Dr. Bennett has been an active member of I/ITSEC since 1994 and has served on two conference committees.

Objective Human Performance Measurement in a Distributed Environment: Tomorrow's Needs

Brian T. Schreiber, Eric A. Watz
Lockheed Martin
Mesa, AZ
Brian.Schreiber@mesa.afmc.af.mil,
Eric.Watz@mesa.afmc.af.mil

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

BENEFITS OF NETWORKED SIMULATION

Networked simulation environments have recently received a great deal of attention. Prior to 1998, warfighters trained complex tactical missions primarily during infrequent range exercises such as the U.S. Air Force's Red Flag. However, even realistic range exercises pose training restrictions. In Red Flag, as just one example, there are space and altitude restrictions. In addition to being costly, resource availability limits the potential number of aircraft. With the advent of distributed mission operations (DMO) multiplayer networked simulations, resource issues and tactical employments are less restricted, thereby allowing warfighters more opportunities to train for wartime requirements of today or possibly even to train to the potential wartime requirements of tomorrow.

Multiplayer simulation allows combat pilots, for example, to routinely train as a four-ship or as part of a larger package of sensor and shooter teams against a varied number and type of threats. In just a single hour-long simulator session, a flight of F-15 or F-16 pilots can experience roughly four to nine individual air-to-air engagements, encounter 20-30 threats or more, and employ dozens of shots. This focused practice can quickly improve a warfighter's Mission Essential Competencies (MECS; see Colegrove and Alliger, 2002). MECs are the 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. Undoubtedly, DMO training improves warfigher MECs, but by what degree do the skills improve? That is, what is the magnitude of training effectiveness benefit in terms of increased human performance? Demonstrated mission performance improvement by the warfighter is, after all, the primary objective of DMO training environments.

QUANTIFYING WARFIGHTER PERFORMANCE IMPROVEMENT

History provides some confidence that future comprehensive in-simulator and transfer studies using objective data will support current DMO as a valuable warfighter learning experience. A number of early studies investigated either in-simulator learning or transfer effectiveness 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).

However, prior research has only involved simple tasks representative of a small portion of a mission (e.g., manual bomb delivery, one versus one air combat maneuvering, etc.). 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. DMO-specific training effectiveness evaluations are needed.

Ideally, an evaluation of DMO training effectiveness would include operator opinions, in-simulator learning evaluations, and transfer of training evaluations to the actual environment (Bell and Waag, 1998; Waag and Bell, 1997). Popular opinion among the DMO community is that networked simulation training is highly beneficial. While we do not doubt this widespread consensus, empirical evidence quantifying the warfighter in-simulator learning improvement is, at best, scarce and, in its very short history, has been limited largely to opinions or subjective ratings (Bennett, Schreiber, and Andrews, 2002; Crane, Robbins, and Bennett, 2000; Krusmark, Schreiber, and Bennett, in preparation; Waag and Bell, 1997; Waag, Houck, Greschke, and Raspotnik, 1995). Objective, quantifiable metrics used in empirical investigations of performance transfer from multiplayer networked environments to the actual environment appear to be nonexistent. Amazingly, hundreds of millions of dollars are being invested in these networked simulation training environments without objective, quantifiable metrics that provide clear evidence of learning or transfer effectiveness and ultimately return on investment (ROI). In other words, the training effectiveness evaluations for *multiplayer networked simulation environments* have rested upon operator opinions and expert evaluation ratings for in-simulator learning assessments—only a small portion of the necessary data for a proper effectiveness evaluation.

OBJECTIVE MEASUREMENT NEED

For networked simulation, subjective evaluations have been the foundation for large-scale aggregate analyses, and those subjective evaluations have generally supported that learning occurs in the DMO training environments. However, those metrics are not sufficient for training effectiveness evaluations, nor are they particularly useful for quantifying warfighter learning improvement translating directly to mission effectiveness. That is, even well-controlled studies reporting significant improvements in average expert ratings for dozens of pilots on tens of missions for constructs such as “situation awareness” or “tactical employment” still do not inform the warfighters, instructors, or the scientific community what impact those changes have on specific and essential knowledge or skills, mission performance such as changes in kill ratios, bombing errors, missile hit ratios, time spent within various range rings, communication step-overs, precise angles/ranges for clear avenue of fire, etc. An objective measurement system directly assessing mission critical parameters validates and *quantifies* the opinion and subjective rating data, thereby providing the ROI data justifying DMO training expenses.

Identifying further potential complications in relying upon subjective measures for training effectiveness evaluations, Krusmark et. al (in preparation) analyzed subjective data for 32 F-16 teams (148 pilots) who participated in one week DMO training research. Pilots flew between seven and nine one-hour DMO missions (222 total scenarios for the 32 teams). Due to understandable logistical constraints, it is not uncommon, as was the case in Krusmark et. al, for the Subject Matter Expert (SME) raters to be employed by the DMO organization and make their rating judgments while observing the missions in real-time. Even though the authors found that the subjective ratings increased over the course of a week of training, suggesting learning and an improvement in performance, it could be argued that the results were due to a vested interest in successful training and/or the lack of using a blind rater methodology. Furthermore, though the overall trend suggested improvement, the authors employed a multilevel modeling approach and found little systematic variability amongst the 40 individual skill constructs rated. The authors concluded that significant learning almost certainly took place, but that plausible alternative explanations for the rating changes exist and the subjective rating system did not appear sensitive enough to differentiate between specific knowledge or skill constructs.

Objective data has not been neglected in DMO environments. Rather, the intent has focused on immediate warfighter feedback. Several DMO facilities demonstrate impressive debriefing capabilities with the ability to replay and/or show vast quantities of objective data (e.g., Nellis AFB, Mesa, Shaw AFB, etc.). However, these systems are limited in that they only provide state-of-the-art capability to show data for *a given scenario or live-fly event*—that is, focused on debriefing a given mission. Certainly these debrief systems provide extremely useful feedback to the warfighter, but they do not permit quantifying learning as part of a formal training effectiveness evaluation. Given these shortcomings, a measurement system capable of automatically aggregating performance over score of missions and warfighters is required.

Historically, objective simulation effectiveness evaluations resulted from customized measurement applications not well suited for other simulation systems (i.e., specific to the stand alone environments). DIS/HLA standardized protocols allow geographically separated entities to engage one another. If a human performance measurement system could be developed with the same level of standardization, *one* system with the same metrics could be used across DMO sites. The number and types of laboratory studies and field in-simulator assessments and the subsequent leveragability of results would be unprecedented. If that same human performance assessment system could also be employed for live-range activities, imagine the potential for cross-comparison of laboratory, field DMO in-simulator learning, and transfer-of-training evaluations.

PERFORMANCE EFFECTIVENESS TRACKING SYSTEM (PETS)

Taking the first step towards achieving a single standardized human performance assessment ability, an automated, objective measurement system was sought for use specifically with DMO. During 2000-2001, researchers, engineers, and SMEs at the Air Force Research Laboratory, Warfighter Training Research Division (AFRL/HEA), developed a Performance Effectiveness Tracking System (PETS) proof-of-concept. Conceptually, this system functions similar to any other entity on a network, except that the system broadcasts almost no data of its own. PETS resides on the network, “listening” to all the data passed through the network interface unit. The network data traffic is captured by PETS, using the data as inputs to real-time human performance measurement algorithms (see Figure 1). For example, by tracking positions the system automatically calculates how much time an entity has spent within certain range boundaries to adversaries. Descriptive statistics, such as number of scenarios, adversaries encountered, missiles fired and outcome measures such as strikers on target, kill ratios, and missile hit ratios, are automatically recorded and aggregated.

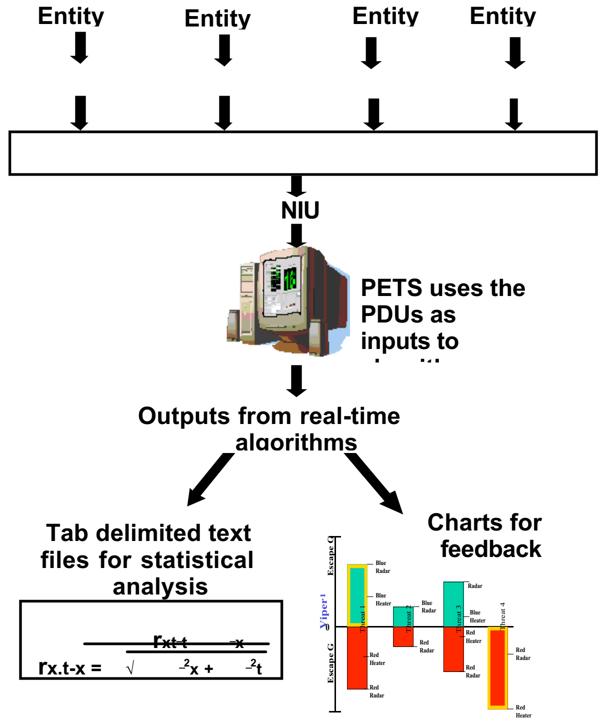

Figure 1. High-level overview of the Performance Effectiveness Tracking System.

PETS is designed to operate in a many vs. many DMO environment and has a number of current capabilities for providing objective human performance metrics within a DMO environment. With these capabilities in mind, the remainder of this paper will describe and discuss preliminary data collected from multiplayer air combat engagements and advocate for increasing the scope of DIS/HLA standards to enable more detailed warfighter performance assessment in the future.

DMO Data Available Today

For the purposes of this paper, a DMO simulation environment can be thought of as having two types of data—external and internal state data. Figure 1 shows how the external state data are shared commonly on the network, therefore being readily available for processing. Simple examples include latitude, longitude, and altitude for every entity. Typically the external state data consist of what is required by network protocol standards (e.g., DIS/HLA), though additional customizations within the standards may be made. Internal state data, however, consist of data typically resident internal to each entity. In Figure 1, the internal state data would remain within each entity, without becoming available to other resources on the network. A simple example would be a switch setting.

The PETS system relies upon the external or "shared" data. From January, 2002 to June, 2003, this system has been used to collect, in real-time, data on over 190 F-16 pilots and over 1,300 scenarios. In real-time, PETS captures over 1,000 variables at 20 Hz, or over one million data points per minute. Most of these data points are either related to entity type or positional information, or can be a calculation from this information (e.g., angles). In addition to outcome measures such as strikers on target and kill ratios, some process measures can and are continually calculated using these external state data. Though over a thousand variables are captured for a variety of research purposes, only a few dozen metrics are currently pertinent for summarizing training effectiveness evaluations. The proof-of-concept system was developed first for use with air combat DMO; a listing of current effectiveness evaluation metrics is provided in Table 1.

Table 1. Types of DMO objective training effectiveness data currently available in the PETS proof-of-concept.

OUTCOME MEASURES	PROCESS MEASURES	MUNITION EMPLOYMENT
Bombers reaching 2nm of target	Time each entity spent within MOR	Who shot what type of weapon at whom
Minimum distance to target bombers achieved	Time each entity spent within MAR	Result (hit, miss, and some types of misses). Also records distance of misses for missiles. For bombs, records left/right & long/short error distances
Fratricides	Time each entity spent within N-pole	2D and 3D range of shot

Mortalities	Chaff/Flare usage	Altitude at pickle
Enemy fighters killed	AIM-9 Clear Avenue of Fire, measured by angles	F-pole for hits and misses
Enemy strikers killed	AMRAAM Clear Avenue of Fire, measured by distances	A-pole
Missiles fired that resulted in a kill	Wingman position in relation to lead	Loft angle at pickle
Scenario demographics (e.g., number of threats presented, etc.)	How often and for how long flight "steps-on" one another when talking on mic	Mach at pickle
Pilot demographics (e.g., flight hours, flight qualification)		G-load at pickle

One type of a process measure linked to the air superiority Mission Essential Competency skill "Controls Intercept Geometry" is the time a pilot violates the heuristic Minimum Abort Range (MAR). This is calculated in the following manner (refer to Figure 2):

1. PETS "listens" to network data, tracking all entities in real-time for force (i.e., red or blue), type (i.e, fighter, bomber, etc.), and position (i.e., latitude, longitude, altitude).
2. PETS then ignores all friendly aircraft.
3. PETS then ignores enemy bomber aircraft.
4. PETS then continuously tracks enemy fighter aircraft position and weapon load.
5. Of aircraft in step #4, PETS continuously calculates each enemy's aspect angle.
6. PETS applies rules: If aspect angle is > 120 degrees (i.e., pointing towards friendly), then given the enemy's altitude, weapon type, and quadrant, is the current range less than that of value in configuration table?
7. If yes, PETS credits the friendly with allowing hostile to penetrate MAR.

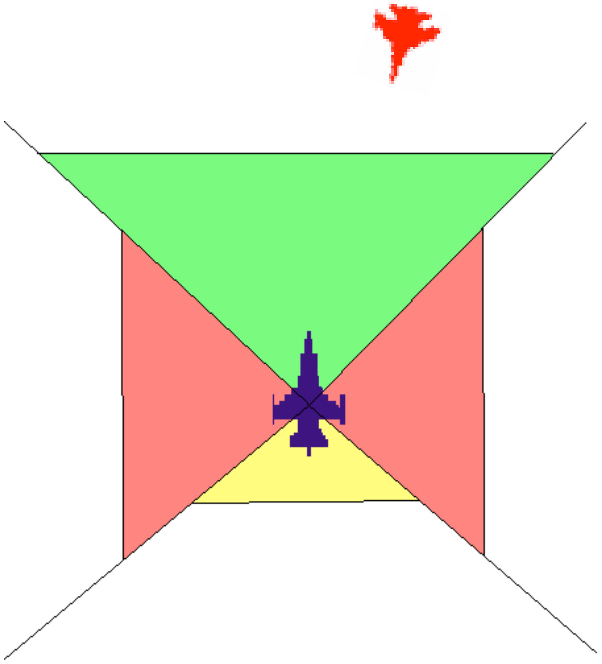

Figure 2. Graphical depiction of the fighter pilot heuristic, MAR. In this example, the hostile aircraft is about to penetrate MAR.

Using data generated from the PETS system, Schreiber, Watz, Bennett, and Portrey (2003) reported substantial objective DMO in-simulator learning for four F-16 four-ship air combat teams (16 pilots). In the current work, we expand upon the results from Schreiber et al. (2003), reporting a much larger data set collected using PETS. Unfortunately, the nature of the air combat data severely limits what we can report. However, the purpose of presenting results here is *not* to report a comprehensive in-simulator DMO learning assessment. Reporting in-simulator results here serves only to illustrate that DMO objective performance is feasible. The primary purpose of this paper is to emphasize and advocate that, by expanding DIS/HLA protocol standards, simulation communities can access even more detailed skill measurements than what is available in Table 1. If the standards expand, the ability for laboratory and field cross-comparison and leveragability of results expands, greatly benefiting the scientific training community, and ultimately, our warfighters.

AN OBJECTIVE DMT/DMO IN-SIMULATOR PERFORMANCE ASSESSMENT OVERVIEW

Participants

During 2002, the PETS proof-of-concept collected data on teams of F-16 pilots who participated in week-long DMO training research events. Here we report data for 19 (95 pilots). All participants were operational United States Air Force F-16 pilots. For the four pilots who flew the pre- and post-test scenarios used in the effectiveness evaluation (76 pilots), all but two were male. The mean age was 32.5 (range 25-48). The mean number of years in service was 10.0. F-16 flight hours ranged from 130 to 2400 hours (mean =850.2). The average team total flight hours was 3400.7.

General Procedure

At AFRL/HEA in Mesa, AZ, F-16 pilots participated in week-long DMO training research. During these weeks, visiting pilots, along with an air weapons controller, flew in a high-fidelity networked 4-ship simulation environment against computer generated adversaries.

Pilots arrived early Monday morning. They completed demographic forms and were assigned anonymous barcode identification numbers. They were then briefed on procedures and objectives before beginning a 3.5-hour mission routine that repeated twice a day throughout the course of the week. The first of nine DMO missions for the week was a familiarity session. Except for the "benchmark" scenarios (discussed later) flown during missions two and nine, the missions followed a building-block training approach. That is, the missions were progressively more complicated as the week progressed.

The 3.5-hour mission routine started with a required briefing given by the flight lead. An hour was provided for the mission briefings. At the completion of the briefing and before "flying" in the simulators, the SME data collector scanned-in the identification number for each pilot/cockpit assignment into the PETS system, thereby providing the automatic link for demographic information used in subsequent analyses.

After briefing, pilots flew DMO air combat engagements for an hour. For any given DMO mission, the four pilots typically flew between four and eight air-to-air combat engagements of the same mission genre while the PETS system automatically recorded all measurements for that team. Each "engagement" or scenario within a mission was a composition where friendly and adversary forces were initialized in the air at greater than 40 nautical miles. After initialization, each scenario continued until either the learning objectives were met or all friendly or all adversary forces were eliminated.

After completing the hour of flying, the pilots were provided 1.5 hours of digital debrief facility time where each engagement could be replayed for the pilots to discuss their performance. The debrief facility included a God's eye view and four avionic displays from each cockpit. This 3.5-hour mission cycle repeated each morning and afternoon until the pilots completed training research around midday on Friday.

For our effectiveness evaluation, participants flew equal but high complexity point defense benchmark scenarios on Monday afternoon (mission two) as a pre-test. The benchmarks consisted of 4 vs 6 adversary fighters with 2 additional bombers; an example is shown in Figure 3. Serving as the post-test assessment, on Friday (mission nine) after completion of the week-long DMO training, pilots flew the mirror-image of their pre-test benchmarks. They flew the post-test scenarios in the same cockpit assignments as the Monday pre-test (pilots were not informed of the mirror-image assessment, nor did any pilot report recognizing the scenarios). All five benchmarks developed were designed to be equally complex. According to results of a scenario complexity analysis employed using the methodology from Denning, Bennett, and Crane (2002), all five benchmarks and their mirror-images were indeed determined to be equally complex.

Over the course of the DMO training week, the 19 teams each underwent nine 3.5-hour mission cycles. For missions two through nine inclusive, each team flew an average of 35 total scenarios against 293 total threats and employed 483 total shots (5,562 total threats and 9,186 total shots for all 19 teams).

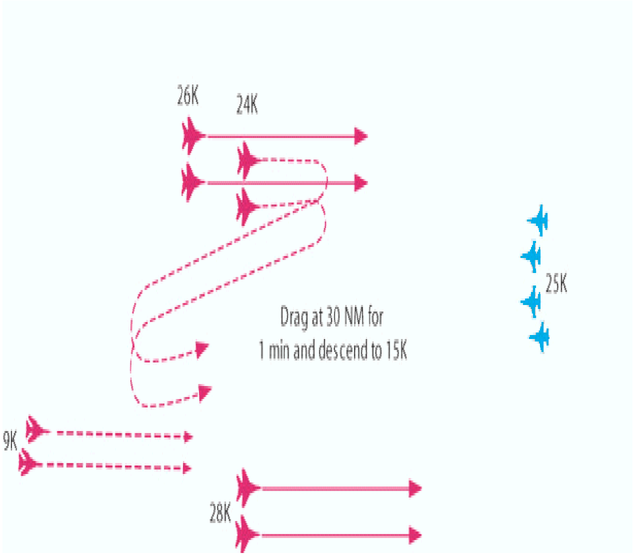

Figure 3. An example point defense benchmark scenario.

Since the objective data is captured, computed, and aggregated in real-time, operator feedback was possible. A statistical batch script file was written and, beginning in October 2002, participants received feedback as to their in-simulator performance improvement. On Friday afternoon, after completing all nine missions, pilots received graphical representation of their performance on many of the objective measures listed in Table 1. Different from traditional one-scenario feedback approaches used in debriefing facilities, this feedback provided mean observed benchmark performance on a given measure. This mean is compared to the 95% confidence interval for all other F-16 teams that have participated in week-long DMO training events (see Figure 4). This aggregate performance displayed over time presented the pilots with feedback never seen before in multiplayer simulation networks. Pilots see their average absolute observed performance before and after DMO training and therefore gain insight into how their performance and behavior changed as a function of DMO training. Furthermore, pilots witnessed their performance in relative terms to the population of F-16 teams that participated in DMO training, gleaning how their in-simulator learning rates compared to other teams representing the F-16 community.

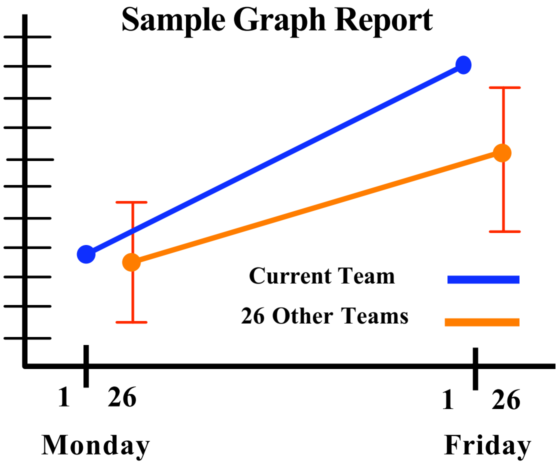

Figure 4. Generic graphical feedback report shown at the completion of DMO. Graphs show current group compared to 95% confidence interval of all previous teams on a given metric (in this example, 26 other teams). Any measure collected can be displayed on the ordinate.

Results

Due to the nature of air combat data, only a limited set of descriptive statistics will be reported in terms of percentage change. Comparing pre- and post-test benchmark performance, the 19 teams (76 pilots), on average, allowed 63% fewer enemy bombers to reach their target, killed 24% more enemy fighter aircraft, allowed 68% fewer F-16 mortalities, and increased the proportion of F-16 missiles resulting in a kill by 7%. Furthermore, other measures suggest that these improvements were not the result of simply increasing a risk bias. The F-16 pilots launched their radar missiles at 8% longer ranges, they survived more frequently, they spent 63% less time within critical ranges to threat fighter aircraft, and they reduced the proportions of threat missiles resulting in a kill by 62%.

DISCUSSION

The limited in-simulator learning data disclosed here represents a robust capability to capture objective human performance data from a DMO environment. Certainly, though detailed analyses are reserved for an alternative reporting system, these high level in-simulator results suggest that a standardized human performance measurement tool and metrics is possible, justifying the desire for its use within the DMO community.

The results suggest that the demonstrated performance improvement was a function of significant learning in the DMO environment. These results—quantifying the magnitude of in-simulator learning improvement--also provide evidence for reaffirming the subjective studies discussed earlier (Bennett et. al, 2002; Crane et. al, 2000; Krusmark et. al, in preparation; Waag et. al, 1995). Furthermore, these objective results show the F-16 teams were not simply sacrificing performance in one area to improve performance in another area, but rather that they were improving in both offensive and defensive skills. By the end of the training week, F-16 teams spent less time within critical threat ranges, they increased weapons employment effectiveness, and their kill ratios increased. In addition to learning other critical skills, we postulate that the F-16 pilots learned where and when to best position their weapons systems in specific inter-aircraft geometric positions such that they could effectively employ their radar missiles, but still avoid vulnerable exposure to the threats' weapons engagement zones.

Efforts are currently underway to enable and test the PETS measurement capability at a sample of other DMO sites. Realizing this next objective would immediately yield benefits such as graphs similar to Figure 4 for the Warfighters and their instructors. For the scientific community, implementation of a standardized DMO human performance assessment capability would allow the cross-comparison of DMO laboratory and field site training results. That kind of leverage allows the DMO community as a whole to collaboratively evaluate different training strategies, impacts of new technologies, Mission Essential Competency skill decay rates, or any of a number of other investigations.

As a final, long-term objective, the same standardized measurement capability and associated human performance measurements could theoretically be used in range exercises. At the Nellis range, for example, much of the data for the aircraft participating in live-fly exercises is passed in a similar manner to current network protocol standards. If a standardized measurement tool with a standardized suite of metrics can be employed at not only the DMO simulation facilities, but also at the live exercise ranges, the scientific potential for discovering the best uses of DMO cannot be overemphasized; objective, in-simulator learning assessments would become *routine* and *any* systematic change within or between similar DMO environments could then be objectively assessed. Furthermore, straightforward transfer of training assessments from the DMO environment to the range becomes possible.

A number of the measures currently available for effectiveness evaluations are outcome related with only a limited number of more process- or skill-oriented measures. While these objective data currently available are highly desirable (i.e., the measures listed in Table 1), a large portion of the 37 MEC air superiority skills (see Colegrove and Alliger, 2002) still cannot be objectively addressed. Therefore, additional measurements, which potentially include a synthesis of both objectively and subjectively obtained data, would potentially permit assessment of more skills more completely and would expand the possibilities for specific MEC skill evaluation, pilot/instructor feedback, and scientific evaluations. However, to expand the objective measurement capability to do this, more *internal* state data need to be available on the DIS/HLA network. At our DMO site, we have made customized enhancements within allowable DIS standards to accommodate PETS' needs. Ideally, any necessary alterations should be identified and *standardized*—not custom—to allow for standardized human performance assessment.

As a simple example, weapons load for an entity is not mandated by DIS/HLA protocol standards and, to accurately measure human performance, our DIS environment provides weapons load in a PDU packet. The MAR measurement described here requires knowing the type of weapon(s) carried by the threat. Without knowing the exact weapons load, the PETS software is forced to assume a "default" weapons load for an entity. For example, if a threat entity carries a "Charlie" weapon but our default assumes the more typical and likely "Alamo/Archer" weapon load, PETS will use incorrect values when calculating the MAR measurement. In proposing and advocating for changes to DIS/HLA standards specifically to include data necessary for performance assessments, the threat entity would have to provide its weapons load at the simulation start, as well as provide an update when each munition is fired. Other examples of internal state variables include fuel load, radar modes, and throttle position, all of which are not currently available through "normal" DIS/HLA simulations.

We propose to extend the amount of standardized external data available on the network by using existing capabilities of the DIS/HLA standards. Upon creation, an entity should send an immediate update of all additional data properties it will expose. During the simulation exercise, all entities exposing this data should be required to send updates when their internal data changes. For example, in a DIS simulation the new data could be passed as a DIS DATA_RECORD object. Discrete data should be updated by the entity whenever the value changes. For example, an update for the "radar mode" property should be sent whenever a pilot switches radar modes. Updates to the weapons load should be sent after each missile launch or gun burst. Continuous data, such as fuel burn, should be updated at periodic intervals and should include both the fuel amount and the current burn rate. The interval at which the data is updated should be determined by a timeout value and/or whenever the rate of change for the data exceeds a specified threshold value.

In HLA networks, the information should be exposed as additional public properties of an entity, and should be made available using standard HLA requests. The updated HLA properties should be broadcast according to the criteria set for DIS updates.

Although we are still developing some of the performance measurements at our installation, we have implemented additional data such as weapons load and radar lock information into the network using the techniques described above. In future enhancements to the DIS standard, we hope to see this additional information incorporated directly into the DIS ENTITY_RECORD objects.

Without variables such as weapons load, fuel load, or other critical metrics being available on the network in a similar fashion to the external state data shown in Figure 1, opportunities for generating new human performance skill measurements are limited. This also then limits the potential for cross-comparison of skill results from DMO laboratory and field sites. Increasing the scope of data available on a network, specifically developing standards within DIS/HLA, alleviates shortcomings in providing the data necessary for more extensive human performance assessments.

The PETS proof-of-concept early successes have generated interest for possible use in other DMO communities outside of AFRL/HEA, including Shaw AFB (F-16s), the United Kingdom (Jaguars and Tornados), the future JSF and F/A-22 community, and the Navy Aviation Simulation Master Plan. As the interest grows and the breadth of possible applications expands, the collective voice advocating for human performance data standards within DIS/HLA protocols also grows.

ACKNOWLEDGEMENTS

This research was performed under Air Force contract #F41624-97-D-5000. The authors wish to thank the contributions of Brett Rigsby, Antoinette Portrey, Margaret Sallay, and Bill Stock for their efforts in supporting the development of this paper. The views expressed herein are those of the authors and do not necessarily reflect the views of Lockheed Martin or the Air Force.

REFERENCES

Bell, H.H. & Waag, W. (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., Schreiber, B.T., & Andrews, D.H. (2002). Developing Competency-Based Methods for Near-Real-Time Air Combat Problem Solving Assessment. *Computers in Human Behavior, 18,* 773-782.

Colegrove, C.M. & Alliger, G.M. (2002). Mission Essential Competencies: Defining Combat Mission Requirements in a Novel Way. Paper presented at the *NATO SAS-038 Working Group Meeting*, Brussels, Belgium.

Crane, P., Robbins, R., & Bennett, W. (2000). Using Distributed Mission Training to Augment Flight Lead Upgrade Training. In *Proceedings of the Interservice/Industry Training Systems and Education Conference,* Orlando, FL: National Security Industrial Association.

Denning, T., Bennett, W., & Crane, P. (2002). Mission Complexity Scoring in Distributed Mission Training. In, *Proceedings of Industry/Interservice Training Systems Conference*, Orlando, FL: National Security Industrial Association.

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 (AFHRL-TR-80-20, AD A094 608). Williams Air Force Base, 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 (AFHRL-TR-77-29, AD A043 648). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.

Hagin, W.V., Dural, E., & Prophet, W.W. (1979). Transfer of Training Effectiveness Evaluation: US Navy Device 2B35 (Seville Research Corporation Rep. No. TR79-06). Pensacola, FL: Chief of Naval Education and Training.

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, D.C.: National Security Industrial Association.

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., Schreiber, B.T., & Bennett, W. (in preparation). Measurement and Analysis of F-16 4-ship Team Performance in a Simulated Distributed Mission Training Environment. Manuscript in preparation.

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., 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,* 87-100.

McGuiness, 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, D.C.: National Security Industrial Association.

Payne, T.A., Hirsch, D.L., Semple, C.A., Farmer, J.R., Spring, W.G., Sanders, M.S., Wimer, C.A., Carter, V.E., & Hu, A. (1976). Experiments to Evaluate Advanced Flight Simulation in Air Combat Pilot Training: Volume I. 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 Air Force Base, NV: U.S. Air Force Tactical Fighter Weapons Center.

Schreiber, B.T., Watz, A., Bennett, W., & 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.

Waag, W.L. (1991). The Value of Air Combat Simulation: Strong Opinions but Little Evidence. *Royal Aeronautical Society Flight Simulator Symposium*, London, England.

Waag, W.L. & Bell, H.H. (1997). Estimating the Training Effectiveness of Interactive Air Combat Simulation (AL/HR-TP-1996-0039). Armstrong Laboratory, AZ: Aircrew Training Research Division.

Waag, W.L., Houck, M.R., Greschke, D.A., & Raspotnik, W.B. (1995). Use of Multiship Simulation as a Tool for Measuring and Training Situation Awareness. In AGARD Conference Proceedings 575 *Situation Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD-CP-575). Neuilly-Sur-Seine, France: Advisory Group for Aerospace Research & Development.

Wiekhorst, L.A. (1987). Contract Ground-Based Training Evaluation. Executive Summary. Langley Air Force Base, VA: Tactical Air Command.

Wiekhorst, L.A. & Killion, T.H. (1986). Transfer of Electronic Combat Skills from a Flight Simulator to the Aircraft (AFHRL-TR-86-45, AD C040 549). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Division.