Integrating and Presenting Performance Information in Simulation-Based Air Warfare Scenarios

Paul H. Radtke & Joan H. Johnston
Naval Air Warfare Center Training Systems Division
Orlando, FL
paul.radtke@navy.mil, joan.johnston@navy.mil

Elizabeth Biddle
The Boeing Company, Training Systems & Services
Orlando, FL
elizabeth.m.biddle@boeing.com

Thomas F. Carolan
Alion Science & Technology
MA&D Operation
Boulder, Colorado
tcarolan@alionscience.com

ABSTRACT

Simulation-based tactical training exercises are ideal settings in which to evaluate performance. The capability to record the second-by-second behavior of participants, the state of supporting equipment, and the location of entities in the problem provides an opportunity to verify team and individual proficiency, and to identify root cause of substandard performance. However, responsibility for determining cause and effect in tactical scenarios is typically left to the expert instructor. In dynamic, fast-paced warfare areas, such as air-to-air combat, the burden on the unaided expert instructor to monitor, record, and assess the interactions and circumstances that determine mission success, is substantial. This is an area where appropriate technology might help the instructor to improve the evaluation of performance.

The Debriefing Distributed Simulation Based Exercises project (DDSBE), an ONR-sponsored 6.3 research and development project, tested alternative technologies for collecting and integrating performance information to aid in the preparation and delivery of post-scenario after action reviews (AARs). The project's objective was to provide the information that instructors need, when needed, in a form that supports rapid evaluation. This paper presents a comparison of different performance data collection, analysis, and debriefing systems, and the performance information they make available to instructors in the context of two distributed training research systems. The first system, built to support the DDSBE research effort, analyzed the performance of two E-2C Naval Flight Officers (NFOs) and F/A-18 Sweep Lead during an air-to-air engagement. Human observers and an automated data collection component collected performance data. The second system, a two-ship F/A-18 simulation built to support training research by The Boeing Company, collected and analyzed performance data for tasks performed by the Escort Lead and Strike Lead during an engagement. The paper presents and compares methods for integrating and presenting the multiple streams of performance information available to the instructor.

ABOUT THE AUTHORS

Paul Radtke is a Research Psychologist with the Naval Air Warfare Center Training Systems Division, Orlando, FL. He holds a BA in Political Science from Western Illinois University and completed the MA in Political Science at Northern Illinois University. Before coming to NAWCTSD in 1994, he served as a Personnel Research Psychologist at the Navy Personnel Research and Development Center in San Diego, CA.

Joan Johnston, Ph.D. is a Senior Research Psychologist and a NAVAIR Associate Fellow at Naval Air Warfare Center Training Systems Division, Orlando, FL. She is responsible for managing basic, applied, advanced technology development, and prototype training research. Dr. Johnston's technical research areas are tactical decision making under stress, team performance and team training technologies, and distributed simulation-based training. She received her M.A. and Ph.D. in Industrial and Organizational Psychology from the University of South Florida.

Elizabeth Biddle, Ph.D. is a Manager with Boeing Training Systems & Services. She has served as Principal Investigator for advanced instructional and training research and development projects. Dr. Biddle earned a Ph.D. in Industrial Engineering and Management Systems from the University of Central Florida in 2001.

Tom Carolan, Ph.D. is a Program Manager with Alion Science and Technology, MA&D Operation. He received his Ph.D. in Experimental Psychology from the University of Connecticut. He has been involved in research and development related to training systems, performance measurement, and human performance modeling for the past 19 years.

Integrating and Presenting Performance Information in Simulation-Based Air Warfare Scenarios

Paul H. Radtke & Joan H. Johnston
Naval Air Warfare Center Training Systems Division
Orlando, Florida
paul.radtke@navy.mil
joan.johnston@navy.mil

Elizabeth Biddle
The Boeing Company,
Training Systems & Services
Orlando, FL
elizabeth.m.biddle@boeing.com

Thomas F. Carolan
Alion Science & Technology
MA&D Operation
Boulder, Colorado
tcarolan@alionscience.com

INTRODUCTION

Recent advances in modeling and simulation (M&S) have greatly expanded the opportunity to conduct multi-platform distributed simulation-based training exercises. For example, advances in M&S interoperability permit Navy Fleet Synthetic Training-Joint (FST-J) exercises to be conducted more quickly, and at significantly lower cost. In March 2006, US Navy, Air Force, Army, and coalition partners participated in a 72 hour FST-J exercise that would have taken over two weeks to conduct just three years ago (Glassburn, 2006). The Navy plans to increase the frequency of such exercises (Jean, 2006). However, this increased demand also results in an increased demand for evaluators who can deliver accurate estimates of mission readiness. Currently, this is a labor intensive (and costly) activity because simulators typically lack embedded tools for automated human performance assessment, diagnosis, and debrief/after action review (AAR). In dynamic, fast-paced warfare areas, such as air-to-air combat, the burden on the expert instructor is substantial. The instructor must monitor, record, and assess the actions and interactions of a large group of performers working on a rapidly changing problem, in which even small mistakes can determine mission success or failure. These tasks are made more complex and time consuming during distributed mission training exercises, in which many teams across different platforms train together but with no face-to-face interactions between instructors and training teams (Neville, Fowlkes, Milham, Merket, Bergondy, Walwanis, & Strini, 2001).

Improving the embedded assessment capabilities of distributed simulation-based training was the focus of an Office of Naval Research (ONR) sponsored program titled "Debriefing Distributed Simulation-Based Exercises" (DDSBE; Johnston, Radtke, Van Duyne, Stretton, Freeman, & Bilazarian, 2004). The DDSBE program developed M&S technologies that can mitigate the added workload of obtaining mission readiness assessments based on objective assessments of combat team and multi-team performance. Technologies were developed that record the moment-by-moment actions of team members, the state of supporting equipment, the location of entities in the problem to verify team and individual proficiency, and the root causes of substandard performance. The DDSBE program tested alternative technologies for collecting and integrating team performance information to aid in the preparation and delivery of post-scenario AARs. The project's objective was to provide the information that instructors need, when needed, in a form that supports rapid evaluation.

The purpose of this paper is to present and compare methods for integrating and presenting multiple streams of performance information available to the instructor. In this paper we compare strategies for performance data collection, analysis, and debriefing systems, and the performance information they make available to instructors in the context of two different distributed training systems. The first system, built to support the DDSBE research effort, analyzed the performance of the E-2C Naval Flight Officers (NFOs) and the F/A-18 Sweep Lead during an air-to-air engagement. Performance data was collected by human observers and an automated data collection component. The second system was built to augment the DDSBE research with a focus on the data collection, analysis, and presentation of tasks performed by the F/A-18 team, comprised of the Escort Lead and Strike Lead, during the air-to-air engagement.

BACKGROUND

The data collected in the two experiments focused on human performance during a simulated air-to-air fighter engagement in a naval strike mission. The two data collection efforts focused on different aspects of the air-to-air engagement, but each followed the same event sequence and tactical context.

An air-to-air engagement consists of a series of voice communications, equipment manipulations, and decisions, performed by individuals or the team, and arrayed along a timeline. Satisfactory performance means performing certain procedures at the correct

time, geometry, and range; using the equipment and systems effectively; making required decisions; and providing necessary information to the right person, accurately, in the prescribed format, when appropriate. The following is a description of the phases and tasks in a generic air-to-air engagement that were used to construct the scenarios, the scripted performance of the trainees used in the studies, and the associated performance measures.

For the purpose of this research, the air-to-air engagement was divided into distinct phases. The pre-commit phase began with the detection of a new, previously unidentified, aircraft by the E-2C command and control aircraft team. Based on the characteristics of the new contact – referred to as a "track" – the E-2C team was expected to assign an appropriate identification designation in the tactical data link and issues a voice report of the contact to the strike package and higher authorities. The fighter element was not expected to take any action regarding the new track except to acknowledge the communication. The fighters relied on the E-2C team to alert them when the contact becomes tactically significant. The pre-commit phase ended when the track's characteristics caused it to be designated as "hostile" and to require a response. The new designation was to be entered in the tactical data link and declared in a voice communication to the strike package.

The "hostile declaration" began the intercept phase. The E-2C vectored the escort to intercept the track. When the fighters acquired radar contact, the E-2C verified that the fighters' contact was the "bandit" in question, based on its reported altitude, bearing, and range from the fighters. The E-2C was then expected to recommend that the fighters "commit" to engage the hostile track. This began the commit phase.

During the commit phase, the E-2C monitored the engagement and passed new information to the fighters, such as any hostile aircraft maneuvers. Otherwise, the E-2C was expected to be silent and not divert the attention of the fighters as they focused on the coming engagement.

During the weapons engagement phase, the fighters attempted to hold the hostile tracks on their radar, while sorting out and targeting the tracks. They also determined the range at which they should release their weapons to minimize their vulnerability to the hostile aircrafts' weapons. The fighter pilots were expected to announce the launches with a voice communication to the E-2C.

The launch of weapons started the merge phase, during which the fighters continued to close the distance to the hostile aircraft, guided the flight of their missile, and watched for an indication that the hostile aircraft had launched a missile against them. The fighters were expected to maneuver to minimize the rate of closure while maintaining radar contact on their target until their missile could automatically track and intercept the hostile aircraft. The pilots were expected to announce this with a voice call to the E-2C. Unless the fighters were obliged to take evasive action to defeat a weapon launched at them from the hostile aircraft, the fighters continued to merge until they observed the destruction of the hostile aircraft, or confirmed that it had survived the engagement. During this phase, the E-2C operator was expected to monitor the engagement and the merge and only communicate with the fighters if there was an immediate threat.

During the post-merge phase, the fighters reported the outcome of the engagement. The E-2C provided an updated picture call to the fighters as they regrouped, prepared to reengage, or returned to their planned route. The E-2C then passed on an engagement report to the rest of the strike package and the Air Warfare commander.

DDSBE SYSTEM

The DDSBE data collection, analysis, and debrief system was developed to support an experiment focused on E-2C - F/A-18 teamwork and taskwork. This system was integrated with a simulation test bed consisting of three positions within a naval strike mission "package". Two of the positions were located on an E-2C command and control aircraft, the Air Control Officer (ACO), and the Combat Information Control Officer (CICO). These two NFOs provide information and coordination to the other members of the mission. The third position was the Lead pilot of the F/A-18 fighter escort, or "Sweep" element, which protects the strike mission from air threats. Because the intent was to test the validity and reliability of the DDSBE system, data collection focused on the pre-scripted individual and team-level performance of the three positions. Team-level performance included the within-platform performance of the ACO and CICO, and the cross-platform teamwork of the ACO and the F/A-18 Sweep Lead.

Four scenario runs were conducted, each containing two air-to-air engagements. The first engagement involved two hostile aircraft, and the second involved a single hostile aircraft, encountering the Sweep Lead and Wingman. During two of the four scenario runs the ACO, CICO, and Sweep Lead followed pre-scripted behaviors to perform at a "nearly perfect" level. During the remaining two runs the trainees performed at a scripted "less-than-satisfactory" level.

The DDSBE performance measurement plan implemented the Event-Based Approach to Training (EBAT; Fowlkes, Dwyer, Oser, & Salas, 1998), which focuses measurement on specific, pre-identified, critical events. When these events are triggered, the participants are expected to perform particular tasks that, in turn, require that they demonstrate targeted skills, knowledge, or other types of competence. This focused approach is based on a sampling of performance and excludes analysis of events not designated to be critical.

Performance measurement relied on both automated data collection and manual input by an instructor. The Virtual Communications Assessment Tool (VCAT), a hand-held device, was used by instructors to record their observations. Two instructors observed the trainees' performance – one assigned to record the ACO and CICO, and the other assigned to observe the F/A-18 Sweep Lead. The hand-held VCAT device warned the instructor when a key or critical event was about to occur and prompted the instructor to record specific observations during the event. The information collected by the human and automated systems filled measurement "slots" within an event-level template of expected actions and indicators. Automatic Performance Assessment (APA) software then compared the observed behavior of the participants with the actions that would be expected by a qualified performer (Carolan, Bilazarian, and Nguyen, 2005). The APA system recorded differences between the observed and expected values for each measurement "slot" in the template, and assigned a numeric score accordingly. The DDSBE system also recorded the trainees' audio communications, and automatically captured screen shots of the trainees' tactical displays at ten second intervals. Instructors also could request additional screen captures via the VCAT tool.

Table 1 presents the 28 performance measurement data items collected by the DDSBE system for each air-to-air engagement. The measures are listed in chronological order and grouped by engagement phase. Eleven of the measures were collected by the automated data collection system that recorded the ACO's and CICO's keystrokes and mouse clicks and the Sweep Lead's control stick movements and button presses.

Table 1. DDSBE Automated and Manual Performance Measures Collected During Air-to-Air Engagements, by Engagement Phase and Event.

Phase	Performance Measures	Automated	Manual
Pre-Commit	ACO "hooks" the new unknown track	√	
	ACO changes track ID to "Unknown Assumed Friendly"	√	
	ACO makes internal "New Track" voice report to CICO		√
	CICO "hooks" the new track	√	
	CICO changes track ID to "Unknown"	√	
	CICO enters the new track information into the tactical data link	√	
	CICO makes external "New Track" voice report to AW		√
	ACO makes external "Picture" call to Strike Package, including Sweep		√
	CICO makes internal "Aircraft Activity" voice report to ACO		√
	CICO "hooks" the track	√	
	CICO changes the track ID to "Hostile"	√	
	CICO enters the new track information into the tactical data link	√	
	ACO makes external "Picture" call to Strike Package, including Sweep		√
	SWL confirms contact report		√
	ACO recommends "Commit"		√
Commit	SWL reports "Commit"		√
	ACO "hooks" hostile track (primary hook)	√	
	ACO "hooks" Sweep lead track (secondary hook)	√	
	ACO makes internal voice report of Sweep "Commit" to CICO		√
	CICO makes external "Commit" report to AW		√
Weapon Engagement	SWL launches weapon via stick	√	
	SWL makes external "Shot" call		√
Merge	SWL makes external "Bulldog" call		√
	SWL makes external "Kill" call		√
	ACO makes internal "Kill" report to CICO		√
Post Merge	CICO acknowledges ACO's report		√
	CICO makes external "Kill" report to AW		√
	ACO makes external "Picture" call to Strike Package, including Sweep		√

The remaining 17 measures were manually collected by the instructors using the VCAT device. Observation scores were used to compute four event-level scores that were aggregated with scores on other relevant events to compute scores for the scenario's training objectives. The individual observation scores also were used to compute mastery scores on individual, team-level, and mission-level competency scales. At the end of the scenario the collected performance data, the track position data, and the accompanying audio and visual recordings were compiled by Assessment Integration software and presented to the instructors for preparation of a debrief. Figure 1 presents the interface of the DDSBE AAR preparation and delivery tool.

The DDSBE AAR tool (Freeman, Salter, & Hoch, 2004) was designed to present the performance data aggregated in chronological order at the event level and by scenario training objectives. Individual events were labeled with "traffic light" symbols of green, yellow, or red to indicate the performance score assigned to the trainees on the event. The red and yellow symbols indicated events in which trainees had performed at a less than acceptable level on one or more tasks or steps within the event. The instructors could "drill down" into an event to identify the specific performer (e.g., CICO) and the performance details (e.g., a missed report) that resulted in the team's score on an event.

The AAR tool also permitted instructors to assess performance in the context of the overall strike mission timeline. When an instructor selected an event from the list on the right of the screen, the geographic display to the left automatically presented the location and heading of all tracks at that moment in the scenario. An instructor also could replay the audio communications and the trainees' tactical displays during the event. Thus, an instructor could present both the assessment of the event and the evidence supporting that assessment in the context of the overall situation.

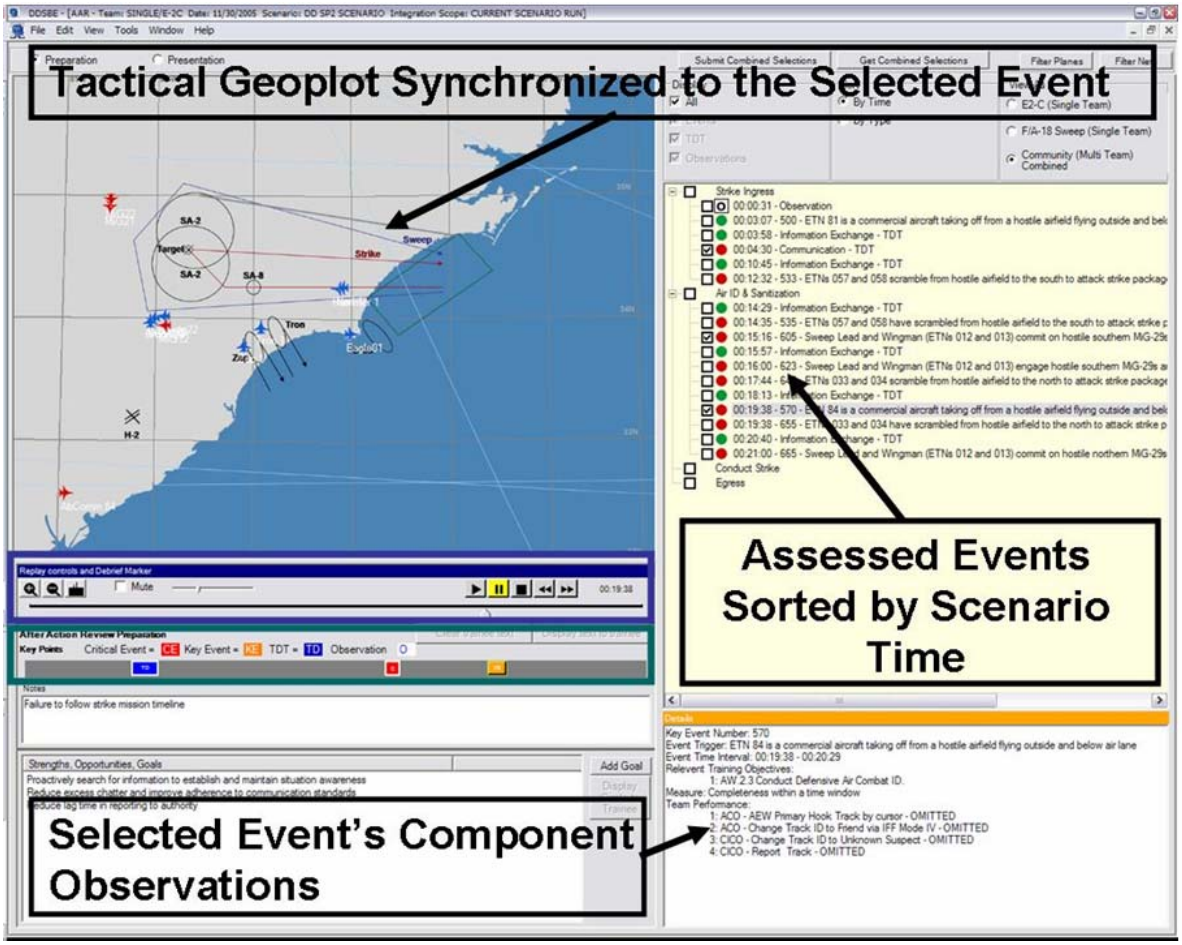

Figure 1. DDSBE AAR Interface

F/A-18 AIRCREW TRAINING RESEARCH

The DDSBE project also developed F/A-18 pilot performance data using virtual and constructive entity position-derived data collected from the distributed network. However, limitations in the simulation environment and project priorities reduced the number that could be tested in the experimental runs described earlier. Therefore, a second research project was initiated through a Cooperative Research and Development Agreement (CRADA) between the Naval Air Warfare Center Training Systems Division (NAWCTSD) and The Boeing Company, Training System & Services (TSS). This complementary project focused on integrating and presenting automated measures of F/A-18 aircrew performance in order to identify strengths and weaknesses in the technologies and provide recommendations for improving the reliability and validity of automated assessment.

The performance assessment test bed was implemented by Boeing TSS and consisted of two F/A-18 unclassified simulators developed by Boeing, a Big Tac™ air threat generator, an Instructor Operator Station (IOS), and automated data collection, analysis, and visualization software. Standard Distributive Interactive Simulation (DIS) network data and non-standard (e.g., button presses, instrument readings) simulation network data were logged with a DIS data logger.

Additional performance data collection was conducted by Alion Science and Technology, MA&D Operation, which was developing a Human-Centered Performance Assessment Tool (HCPAT) under a Small Business Innovative Research Phase II project. The HCPAT research project had developed automated and semi-automated performance measurements to evaluate the F/A-18 aircrew during the engagement and merge phases described earlier. Alion integrated the HCPAT with the Boeing test bed to implement automated and semi-automated metrics for testing. DIS communication middleware was developed as a plug-in to HCPAT to allow the software to observe the network traffic for relevant simulation entity state data; an air combat domain plug-in was created to specify the relevant objects in the performance assessment environment.

The F/A-18 stations were used by the Strike Lead and Wingman roles, and the IOS was used to support an E-2C role-player. The purpose of the E-2C role was to support the information exchanges that are part of the engagement and merge phases, but was not a focus of the performance assessment. The missions were geographically located in the vicinity of Elmendorf United States Air Force Base, Alaska.

Similar to the DDSBE approach, the EBAT methodology was used to fine-tune the scenario and guide the automated and semi-automated measures. A task analysis by Brobst, Geis, and Brown (1999) that organized the performance measures by the F/A-18 aircrew performance elements, air crew skill, and mission phase was leveraged as the basis for organizing the metrics into competencies. A list of scenario events expected during each mission phase was created, and expected tasks and actions were linked to each event. Measures and performance standards were created for a sample of the event tasks and selected for implementation based on mission requirements input from the Subject Matter Experts (SMEs) and the simulators' capabilities. Metrics were designed to generate automatically or with observer input depending on the data available from the flight simulators.

A secondary objective was to identify technical data requirements for constructing specific F/A-18 performance elements. Metrics that only required DIS data could be implemented on a standard DIS network. However, access to non-standard DIS data required software modification. In this experiment, non-standard DIS data was obtained through a Protocol Data Unit (PDU).

Four post- Fleet Replacement Squadron-level air-to-air scenarios, each successively more difficult, were scripted by an F/A-18 Subject Matter Expert (SME). The scenarios involved two F/A-18 pilots (Strike Lead and Escort Lead) and an E-2C role-player (ACO). The experiment was designed to analyze the reliability and validity of the metrics across two performance levels within scenarios of differing difficulty. The four scenarios were each performed by the SMEs three times; once to standard, and twice not to standard.

In the non-standard performance conditions, the F/A-18 pilots deliberately exhibited pre-specified behaviors to test the metric's ability to accurately report the greater variability in performance. Each mission was designed to affect performance on a specific training objective. The missions followed the generic fighter engagement timeline described in the background section. The timeline was adapted to the experimental mission timeline and varied by the complexity of the threat fighters' performance in the four conditions.

Automated air combat measures were developed based on existing air to air combat algorithms developed for the Navy DDSBE project (Carolan, Bilazarian and Nguyen, 2005), analyses and measures developed for

the Air Force (Portrey, Schreiber, & Bennett, 2005) and new algorithms developed for the Boeing aircrew training research environment. The approach was to capture performance relevant data and use the data in metrics to evaluate warfighter performance with respect to higher level training objectives, such as aircrew tasks, mission phases, and underlying competencies. Some of these measures are considered first approximations since not all the data to accurately compute the measure was available. Table 2 presents automated F/A 18 aircrew performance measures developed and tested.

Observer-based measures were constructed for most of the task areas. These were focused on communications - the completeness, timeliness and format of voice reports, adherence to task procedures, and tactical decision making. In addition teamwork measures were also available to the evaluator. These were not event specific measures but provided the opportunity to assess specific aspects of team work observed throughout the scenario.

Table 2. Automated F/A-18 Aircrew Performance Measures

Maintain Mission Timeline
• Time and distance off at waypoints
Weapon Launch
• Range at missile launch • Clear avenue of fire • Tactical advantage – Relative speed and altitude • Acceptable launch region • Crank maneuver
Defensive Maneuvers
• Within E-Pole range/orientation to threat • Escape maneuver executed • Maximum G-force attained • Time to achieve escape range and heading
Maintain Mutual Support
• Outside mutual support range or altitude
• Outside contract speed and altitude value ranges

Integrated Assessment Approach

During the exercise an evaluator using a networked tablet style computer with the HCPAT software observed performance, selected events to assess and entered assessment data. The assessed events were displayed on an event log. The evaluator had the option of entering events and completing the assessments at a later time. The single evaluator assessed between 10 and 20 events during each exercise run. These items included the timeliness and completeness of voice communications, and the quality of tactical decisions.

The automated assessment module monitored the scenario entity state data through the DIS connection, detected performance events, and triggered measures. The performance events and measures were recorded in the event log and made available to the evaluator. An additional alert feature indicating that an event of interest had occurred was still in development and not available during the test.

The automated performance measures were designed to record deviations from expected performance standards, as in Outside Of Mutual Support Range, or a value to be compared against performance standards such as Within E-Pole Range. Automated measures can be event specific measures or global measures monitored as appropriate throughout the scenario. Global measures consisted of detecting and flagging observed deviations from expected performance criteria.

In addition to fully automated measures, which required no human intervention, semi-automated measures were employed to support the observer assessment process. One example is the automated calculation of time between events, where one event is an observer selected voice report. Another measure is the range between entities when a particular event is triggered. Since these are based on evaluator response time, they provide estimates to support the evaluator's assessment.

This is the initial step of the assessment process. The deviations are recoded in the event log and linked to higher level measure categories through the structure of the event tree or through predefined analysis groups. The software supports a number of approaches for using this performance data for assessment and feedback. The first approach uses the automated performance measures to support the evaluator in making assessments. For many of these dynamic measures the assessment can be very context dependent. Flagging potential problem areas and providing the evaluator with performance evidence, behavior anchors, and a rating instrument, allows the evaluator to make the assessment based on observations, context and performance evidence. Simple examples include Maintaining Mutual Support, staying with contracted speed and altitude ranges. We found multiple departures from mutual support range under the 'good' performance conditions. Many were small departures, others were larger, such as, to investigate a potential threat. The evaluator reviews the performance data and makes the judgment on how to assess.

A second approach is to build in automated assessment algorithms that assign a value to a performance instance

or set of instances based on predefined standards and context information. Some of these assessments are built into the measure, such as, a simple pass or fail for clear avenue of fire. Others require triggers to turn measures on and off. In addition, other standards change depending on whether they are performed pre- or post-commit. Others require a more detailed situation assessment and expected performance model, such as assessing targeting decisions.

Real-Time and Post-Event Analysis and Presentation

The *Evaluator* is an automated analysis tool prototype that can be used to create metrics in near real-time during the performance of a training mission and/or on completion of a training mission. We used the post-event evaluation approach in order to create a quantitative, summative value for the measures we collected during the experimental scenario runs. For example, the Maintain Mutual Support metric returns the average range (in nautical miles) between the aircraft over a period of time. If the average distance is within an acceptable maximum range the Maintain Mutual Support value can be further qualified as a "pass." These results can be displayed as "passed" (green) or "fail" (red), based on the raw metric result. Metrics coded yellow, (e.g., Shot Kinematics) involved two quantifiable variables – in this case, the altitude and speed of the aircraft at the time the shot was made. If only one parameter was within standard, the metric evaluated as "partial pass," and displayed with a yellow symbol. The post-event analysis method was used to verify that the SME's performance was assessed as intended.

A complete analysis of the data we collected is still under review. However, initial findings from the experiment enabled us to identify critical weaknesses in the simulation and assessment system that pointed to needed improvements in technologies. Although SMEs had performed to pre-scripted actions, the post-event analysis indicated their actual performance on the scenarios, in many cases, did not match expected performance on the measures. An in-depth analysis of the raw metric data and post-event discussions with SMEs provided valuable insights on the major causes of the inconsistencies in the assessment system results as described in the following.

Mismatch between expected performance and simulation test bed design. Although the performance metrics we developed were specified according to real world F/A-18 pilot behaviors, the simulation test bed lacked some critical functionality in order to be implemented in an unclassified environment that would have allowed the SMEs to perform to expectations. We understood in advance that some of the SME actions would be "artificial" compared to real world behaviors, and as it turned out, the assessment results enabled us to identify this problem.

Task complexities. The parameters used for evaluating performance may have been too constrictive given the complex nature of some of the pilot's tasks. The SME's review of post-event analyses enabled us to understand the extent of the complexities of the performance elements that the metrics were assessing as well as the situation-dependent nature of the metrics.

Accuracy of performance measures algorithm. In some cases the performance measures algorithms did not accurately evaluate the task. The process of evaluating the data and talking to the SMEs enabled these metrics to be refined.

Figure 2 presents a snapshot of sample performance data presented in the realtime *ResultsViewer* display. It is a prototype data visualization tool that is used to display the near, real-time metrics during the performance of a training mission or during a mission playback, such as during an AAR. The real-time *ResultsViewer* approach is to provide the instructor with a graphical display of the metric as it evaluates data in near real-time. This approach can be used during the performance of the training exercise or the *ResultsViewer* can be played back in synchronization with a mission playback during the debrief session. These displays can be used to alert the instructor to a particular situation that may not be detectable through human observation or due to the complexity of the many events that the instructor must simultaneously observe.

The advantages of an integrated assessment approach is it can provide different automated performance data to training evaluators and training participants at different times during or post exercise to support ongoing assessment, diagnosis, and performance feedback needs at different levels of analysis. With a focus on providing formal assessment (ratings) for AAR, one *HCPAT* product is a drill down assessment report implemented as a set of PowerPoint slides. The assessment report displays the color coded ratings and associated comments at each level down to the performance instances. The AAR leader can start at the highest level; for example, the mission phase, or Mission Essential Task level, and then drill down to specific performance instances in the context of the overall scenario situation.

Figure 2. Real-time *ResultsViewer* Display

CONCLUSION

Both the DDSBE and F/A-18 Aircrew Training Research systems provided an opportunity to test and evaluate different approaches to collecting, analyzing, and presenting performance data regarding team and collective performance in a distributed simulation training environment. This type of experiment was critical to identifying the complexities, strengths, and weaknesses of automating assessment of team performance. The following guidelines are based on the results and feedback received during the various experiments.

1. Use the EBAT approach for scenario and performance measurement design: The EBAT approach involves the development of performance measures and data collection requirements during the scenario design process. Therefore, human observation requirements are pre-defined, which will result in minimized workload and simplified data collection processes. This will serve to improve the reliability and validity of the data collected and subsequent assessment. Additionally, the EBAT methodology reduces the tendency to collect data on "everything." Experience has shown that this method does require clear segregation of events that do not influence each other, and occur in the order expected. In order to prevent a reduction in the realism of the scenarios due to these constraints, and to allow for assessment when performers react to events, it is important to develop flexible event-based metrics that can adapt to the context of the scenario in real time (e.g., Biddle, Perrin, Dargue, Lunsford, Pike, and Marvin, 2006).

2. Concentrate performance assessment on known, verified (or verifiable) relationships between observed behavior and likely gaps in certain competencies: Focusing the assessment on known relationships between specific behaviors and gaps in competencies facilitates the diagnosis of root cause. These relationships are now found largely in the experience of SMEs and remain to be captured by training practitioners. Consequently, this diagnostic

process needs to be supplemented with human observation during the event to verify that root cause diagnosis is accurate and not due to an unforeseen event or training system failure.

3. Focus on specific events vice general observations (i.e., "You need to improve your communications!") during debriefs: The use of specific events from the training scenario to discuss an instructional point will improve instructional benefits by providing feedback in context of a specific event. So that expert instructors do not feel excessively controlled by the focus on specific events and specific observations, the post-event automated results, in conjunction with post-event semi-automated results, can be used to provide global observations and evaluations, as long as the instructor can then point to specific events in the scenario.

4. Focus analysis on processes as well as outcomes: The integration of process and outcome measure assists in providing understanding of how team and individual behaviors contribute to event outcomes.

5. Use graphical presentation of performance measures updated in near real-time: Real-time visualization of performance can be used to assist or alert the instructor in diagnosing trainee performance problems and providing real-time feedback or scenario modification. Additionally, the real-time assessment information provides instructors with detailed information regarding student performance that may not be obtained through human monitoring or objective analysis. Real-time visualizations do not provide an overall report on the metric so it should be used in combination with post-event metric results.

6. Balance real-time and post-event automated performance assessment and scoring: A summative, post-event metric provides a quantitative value to provide meaning regarding a "pass" or "fail" evaluation. Real-time ratings may be based on incomplete or premature interpretations of events. The results of both processes need to be considered in conjunction with each other to produce the most accurate and useful feedback to the trainees.

These guidelines are by no means fully-conclusive, and the authors recommend that research continue in this area to enable greater reliability and validity in automating the test and evaluation of training effectiveness. In many cases, the guidelines are more cautionary than prescriptive, which also argues for more thought and testing in this area. The challenge is to integrate and analyze objective performance data from simulation environments so that it is useful for assessment, diagnosis and feedback. This means analyzing both the capabilities of the simulation to support the performance of the tasks being trained or assessed, and the degree to which the data produced in the simulation reflects the trainees' competence to perform those tasks in the real world. It also underscores the importance of empirically testing and validating all aspects of the performance measurement system.

MEMORIUM

We dedicate this paper to the memory of Paul Radtke who passed away during the time it was completed. In his 18 years as a top notch Navy scientist, Paul worked hard to achieve many successes in transitioning scientific products to the research community, the schoolhouses, the operational Navy, and to our joint and coalition partners. He was a great friend, collaborator, and a mentor to all of us, always finding ways to make our work together both effective and fun. We will miss him very much.

ACKNOWLEDGEMENTS

The authors would like to extend thanks to Hugh Carroll (BGI), Steve Dix (Boeing), David Fries (Boeing), Mike McCleod (Thunderbolt), Jeff Miller (Alion), Richard Plumlee (NAWCTSD), LCDR Chris Provan (NAWCTSD), Erick Weber (Alion), and Jake Wigglesworth (Boeing) for their efforts in planning the scenario, specifying performance measures, and participating as role-players during DDSBE and Augmented DDSBE experiments.

REFERENCES

Biddle, E., Perrin, B., Dargue, B., Lunsford, J., Pike, W.Y., & Marvin, D. (2006). Performance-based advancement using SCORM 2004. In the Proceedings of the 2006 Interservice/Industry Simulation, Training, & Education Conference [CD-ROM]. Orlando, FL.

Brobst, W.D., Geis, L.A., & Brown, A.C. (1999). NSAWC aircrew training study: Methodology and analysis (Report No. CRM 98-171). Alexandria, VA: Center for Naval Analyses.

Carolan, T. F., Bilazarian, P., & Nguyen L. (2005). Automated individual, team, and multi-team performance assessment to support debriefing distributed simulation based exercises (DDSBE). In the Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting [CD-ROM], Orlando, FL.

Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-based approach to training (EBAT). International Journal of Aviation Psychology, 8, 209-221.

Freeman, J., Salter, W.J., & Hoch, S. (2004). The users and functions of debriefing in distributed, simulation-based team training. In the Proceedings of the 48th Annual Conference of the Human Factors and Ergonomics Society [CD-ROM], New Orleans, LA.

Glassburn, J. (2006, April). U.S., coalition forces conduct at-sea training without leaving the pier. Navy Newstand Online. Retrieved June 11,2007, from, http://www.news.navy.mil/ search/display.asp?story_id=22943

Jean, G. (2006, September). Navy's virtual training exercises expanding in realism and scope. National Defense Online. Retrieved June 11, 2007, from http://www.nationaldefensemagazine.org/issues/2006/September/NavyVirtual.htm

Johnston, J. H., Radtke, P. H., Van Duyne, L., Stretton, M., Freeman, J., & Bilazarian, P. (2004). Team training in distributed simulation-based exercises. In the Proceedings of the 48th Annual Conference of the Human Factors and Ergonomics Society [CD-ROM], New Orleans, LA.

Neville, K., Fowlkes, J., Milham, L., Merket, D. C., Bergondy, M. L., Walwanis, M., & Strini, T. (2001). Training team integration in a large, distributed, tactical team: A cognitive approach. Proceedings of the 23rd Annual Interservice/Industry Training, Simulation and Education Conference (pp.1035-10), Orlando, FL,

Portrey, A. M., Schreiber, B.T., & Bennett, W., Jr. (2005). The pairwise escape G-metric: A measure for air combat maneuvering performance. In the Proceedings of the 2005 Winter Simulation Conference (1101-08), Orlando, FL.