

Running head: SPOTLITE Method

Draft

Measuring Team Performance in Complex and Dynamic Military Environments:
The SPOTLITE Method

Jean MacMillan

Eileen B. Entin

Rebecca Morley

Aptima, Inc.

Winston Bennett, Jr.

Air Force Research Laboratory

Measuring Team Performance in Complex and Dynamic Military Environments:

The SPOTLITE Method

Abstract

The Scenario-based Performance Observation Tool for Learning in Team Environments (SPOTLITE) is a systematic method for developing team performance measures that are simultaneously linked to learning objectives and to events embedded in the training scenarios. The method was applied to develop a measurement instrument for four-person teams of F-16 pilots training for air-to-air combat in a high-fidelity simulation environment. The instrument is comprised of behaviorally-anchored rating scales that are tied to observable behaviors that tap critical knowledge and skill elements and can be assessed at specific intervals during a training scenario. We implemented the measurement instrument in a handheld computer to support fast and accurate entry of assessment data as a team executes a scenario. An experiment designed to test the measurement instrument demonstrated its sensitivity, reliability, and validity. The SPOTLITE measure-development method is readily extensible to other simulation-based team training environments.

Measuring Team Performance in Complex and Dynamic Military Environments:

The SPOTLITE Method

Introduction

Computer simulations have become an essential technology for training complex dynamic multiperson tasks. Learning that at one time could have taken place only in the real world can now occur in the “virtual world” that is created by realistic simulation environments. These environment can involve multiple individuals performing as a team or as a team of teams. Simulations are being used to train aircraft crews, nuclear power plant operators, military command teams, crisis action teams, medical teams, and many others. The measurement of performance in these complex environments remains a challenge, however. Meaningful quantitative measures of learner performance are needed to provide feedback to learners to make the most effective use of these complex simulations for training.

This paper describes the Scenario-based Performance Observation Tool for Learning in Team Environments (SPOTLITE) method for developing team performance measures.¹ The SPOTLITE method addresses one of the most difficult problems for performance measurement in complex multiperson environments: where to focus the measurement. Typically, a myriad of factors could be measured in a complex team task. The SPOTLITE method focuses the measurement on those factors that should be measured. The goal is to develop feasible, reliable measures that are sensitive to variability in performance, diagnostic of performance difficulties, and validated by their relationship to overall outcomes.

¹ This worked was sponsored by the Air Force Research Laboratory Warfighter Training Research Division.

Performance Measurement for Effective Simulation-Based Training

Providing feedback to learners is widely accepted as a cornerstone of effective training for individuals as well as teams (Brannick, Salas, & Prince, 1997; Salas & Cannon-Bowers, 2001). Simulation-based training environments offer the opportunity for “event-based training” (Johnston, Smith-Jentsch, & Cannon-Bowers, 1997) in which the events in the simulated environment are consciously shaped to create situations for specific types of learning. The simultaneous design of training scenarios that include learning opportunities and the development of performance measures linked to these opportunities provide a tool for the effective use of simulation environments for training. Figure 1 shows this vision for the interlinked development of simulated scenario events and performance measures, with both driven by the objectives for the training.

The SPOTLITE project focused on developing performance measures that were simultaneously linked to events embedded in training scenarios and to learning objectives. The domain of application was a team of four F-16 fighter pilots training to fight as an integrated team (a “four-ship”) in a high-fidelity air-to-air combat simulator. The measurement instrument developed was driven by both top–down objectives for what should be learned by the pilots and bottom–up opportunities that controlled what could be learned (and measured) in the simulated scenarios.

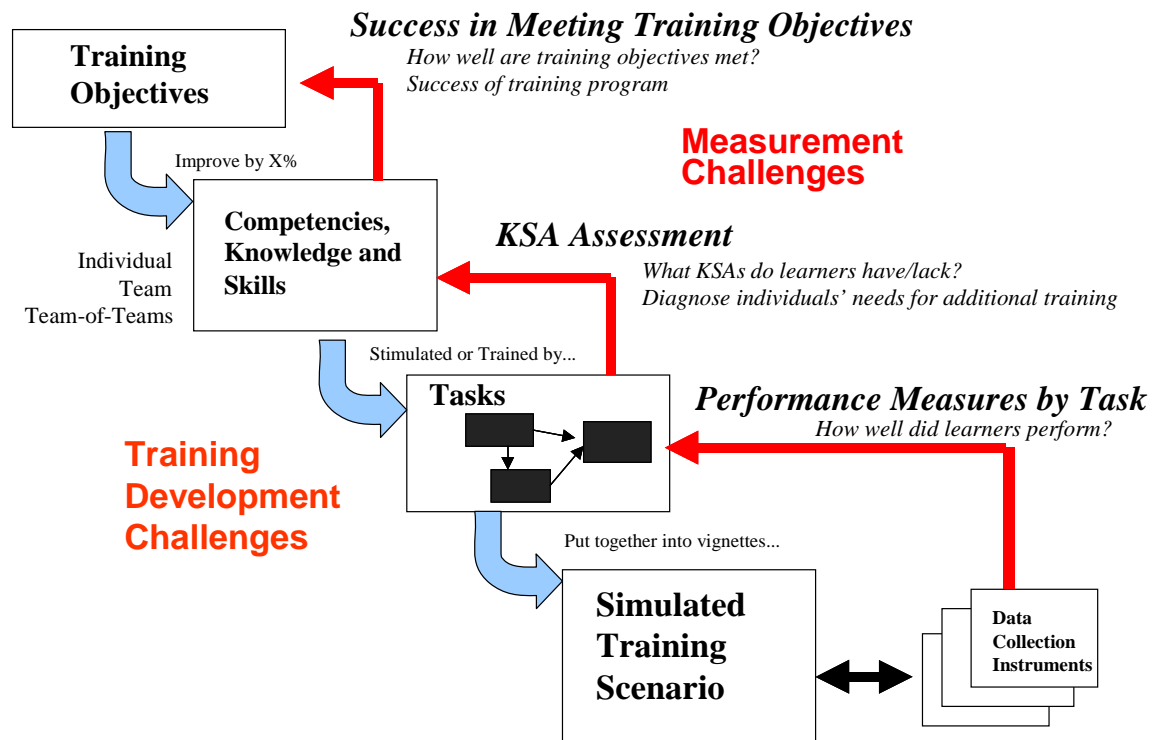


Figure 1. Linking scenario design and performance measurement for effective simulation-based training.

The Challenge of Performance Measurement in Complex Environments

Although the importance of measuring performance in simulation-based training environments is well understood, the design of such measures remains a challenge. Complex multiperson simulations occupy a place in between measurement in tightly controlled laboratory experimentation and measurement in completely uncontrolled real-world situations. In complex multiperson simulations, some degree of control is possible over the initiation of scenario events, but the dynamic actions taken by the multiple players can cause the same scenario to unfold quite differently over time. Simulations thus offer some of the control of the laboratory, but present the real-world problem of defining appropriate measurements.

An advantage of simulation environments is that they typically offer greater opportunity for data collection than the real world. Simulations can potentially record every action taken by every participant over a period of hours or days. Combined with video and audio recording, a simulator-generated database can record and replay every detail of the training session. This ability to record everything can create problems, however, if no methods are in place to filter or condense the data to generate a smaller set of meaningful measures. Recording behaviors is not equivalent to understanding them, and meaningful measurement and feedback often requires human observation and judgment (Baker & Salas, 1997). This necessary reliance on human observers, however, limits the amount of data that feasibly can be collected. A human observer cannot observe everything in a complex simulation simultaneously—attention must be focused on the most useful areas for measurement.

The complex interaction of events and actions in the simulator also presents a measurement challenge. A simulated situation rarely has only one “right answer” or path to success. Simulators typically can provide outcome data for complex tasks, but these outcome data are not necessarily the best basis for effective feedback. As Johnston et al. (1997) pointed out, teams may achieve an acceptable or even a stellar outcome through a stochastic roll of the dice, even though the process they followed was not one that typically would result in success. Teams should not be rewarded for “getting lucky” in the training session. Feedback measures should be based on the best practice behaviors that result in the best outcomes on average, not just in any one instance.

The Focus of SPOTLITE

The goal of the SPOTLITE project was to develop a behaviorally anchored, observer-based measurement instrument for use in an F-16 air combat training environment. The

instrument was to (a) focus on the most important measurement points for assessing and diagnosing learner performance, linked to specific events in the simulated scenarios; (b) be feasible for a single observer to complete during a training session; and (c) provide a sensitive, reliable, valid assessment of the performance on an F-16 four-ship pilot team.

The focus on observer-based measures was driven by the judgment, as discussed above, that the complexity of the simulated scenario precluded a simple mapping of actions to outcomes. Some aspects of performance in the scenario require expert judgment for meaningful measurement and feedback. We anchored the observer-based rating scales to specific behaviors in order to increase their reliability across observers and their usefulness to learners. The F-16 training program simultaneously is developing methods for extracting and condensing data directly from the simulator to be used in conjunction with the observer-based measures (Schreiber, Watz, & Bennett, 2003).

The F-16 Four-Ship Air Combat Training simulation facility at the Air Force Research Lab (AFRL) in Mesa, Arizona, is a high-fidelity simulation that allows F-16 teams to train on multiple defensive and offensive scenarios. The simulation system records the scenarios as teams perform them, including all their cockpit information systems as well as the out-the-window view. Because the simulation facility at Mesa is in almost constant use for operational training, a large body of recorded scenario runs have been accumulated and replay facilities are available. Using these capabilities, we were able to work closely with subject-matter experts (F-16 instructor pilots and training developers) to create the SPOTLITE measurement instrument. The repeatability associated with the simulation, the use of selected benchmark scenarios that begin the same way for all learners, and the replay facilities of the training center were assets in both the development and the evaluation of the measurement instrument.

Team Performance Measurement Approaches

The SPOTLITE measurement approach can be contrasted to other team performance measurement approaches. Differences result from the specific goals for the measurement instrument and the nature of the F-16 training environment.

Construct-based measures of teamwork (Team Dimensional Training). Team Dimensional Training is an event-based approach to assessing measures of team process and outcomes (Smith-Jentsch, Payne, and Johnston, 1996; Smith-Jentsch, Johnston, and Payne (1998)). Using this approach, Smith-Jentsch, et al. (1998) identified four teamwork process dimensions that are important for effective performance. A substantial body of research (Entin & Entin, 2000; MacMillan, Paley, Entin, & Entin, 2004) has established stable, measurable dimensions of team performance that are reliably associated with positive team outcomes. Effective performance on dimensions of team performance such as communication, monitoring of teammates, and back-up behavior (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995) has been shown to result in superior team performance for a variety of team types. Trained observers can reliably rate these dimensions of performance over the course of a scenario. Our goal for the SPOTLITE measurement instrument was more specific—to develop performance rating scales linked to specific events and behaviors in selected training scenarios. For example, rather than having the observer provide one rating at the end of a training session for the quality of communication throughout the session, our goal was to identify and rate specific events in the scenario when communication was likely to be especially critical and difficult. By tying the ratings to specific scenario events and specific behaviors, we hoped to increase both the reliability of the ratings and the diagnostic value of the measures in order to provide more useful feedback to the learners during the After-Action Review that follows each training session.

Distributed team assessment instruments (TARGETS). The TARGETS methodology (Dwyer, Fowlkes, Oser, Salas & Prince, 1997) has been used to develop a performance measurement instrument for aircrew coordination training. TARGETS is especially designed to produce ratings for team performance when teams are distributed across multiple locations. It employs rating scales that are tied to specific scenario and task events and that describe specific concrete behaviors predetermined by subject-matter experts (SMEs) to be appropriate at that point in the scenario. Observers need not be SMEs—the expertise is built into the instrument. Initially, we expected the SPOTLITE instrument to resemble that of TARGETS. Once measure development was underway, however, we found that the intense, fluid, dynamic complexity of an air-combat engagement required a different approach. Although we tied ratings to events in the scenario, we were not able to eliminate the need for a SME to evaluate performance—too many complex, interdependent factors affected the level of performance. These situations have no “right answer,” although some behaviors are certainly more effective than others. Also, we found considerable variability in how the scenarios unfold over time, depending on the actions taken by the pilots. The point of entry, initial flight path, and capabilities of enemy aircraft are prescribed in the scenarios, but their behavior during engagement depends on how the F-16 pilots choose to fight. We therefore tied our ratings to specific periods of the scenario and to specific types of engagements, rather than to predetermined events in the scenario.

SPOTLITE Method for Measure Development

SPOTLITE measure development followed a systematic path as shown in Figure 2, starting with the competencies, knowledge, and skills that F-16 pilots need to acquire to be considered ready for air-to-air combat missions. SPOTLITE linked those competencies to

observable behaviors and selected those observable behaviors that can and should be measured in the training simulation.

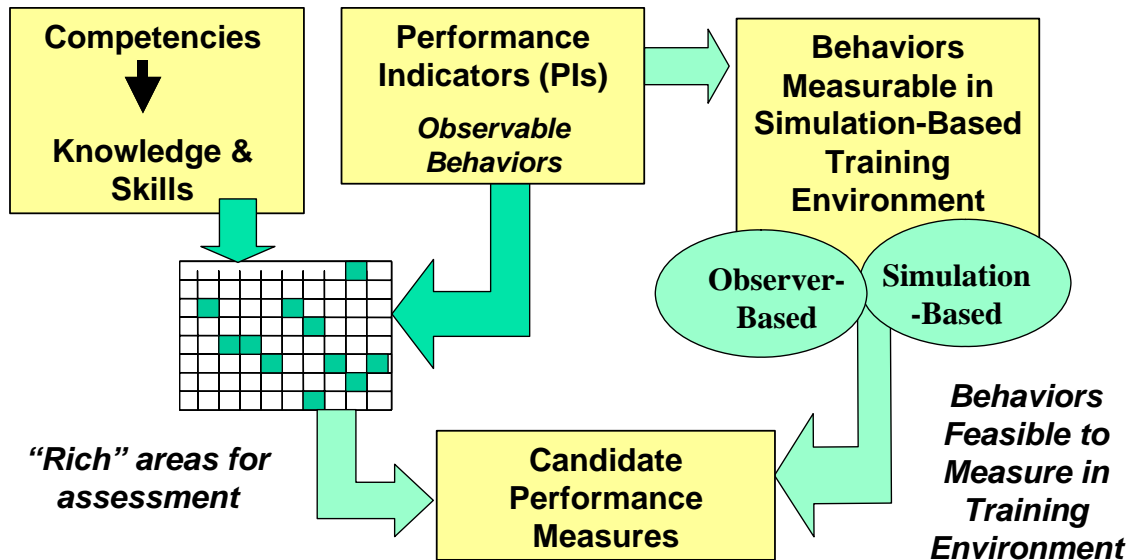


Figure 2. SPOTLITE measure-development process.

Identify Competencies, Knowledge, and Skills Needed for Effective Performance

In developing the SPOTLITE measures, we were able to draw on the products of a parallel effort to identify the Mission Essential Competencies (MECsSM) required for fighter pilots in air-to-air combat missions and the knowledge and skills that support those competencies (Colegrove & Alliger, 2002). These competencies, knowledge, and skills were defined through an iterative series of knowledge elicitation sessions conducted with SMEs—F-15 and F-16 instructor pilots. Overall, this process identified 7 competencies, 12 types of knowledge, and 25 skills. Table 1 shows an example of one of the MECs and some of the knowledge and skills that support that competency. The competencies, knowledge, and skills defined for air-to-air combat constitute the training goals for the simulation environment—the top box in Figure 1—and are therefore the starting point for measure development as shown in Figure 2.

Table 1

Example of a Mission Essential Competency (MEC) and Knowledge and Skills Required for That Competency

MEC (example)	Knowledge	Skills
Intercepts and targets factor groups	Communication standards	Builds picture
	Engage criteria	Integrates sensor output
	Follow-on options	Radar mechanization

Identify Observable Performance Indicators

The next step in measure development was to translate the competencies needed for mission effectiveness into observable behaviors that were candidates for performance measurement. This was done through a two-day workshop with SMEs, focused around specific types of missions. During the workshop, we identified Performance Indicators (PIs) by asking the experts to describe behaviors that they had seen frequently as instructors that indicated to them that a pilot or a team of pilots (a) were at a low level of readiness and needed more experience and instruction or (b) were at a high level of readiness and were ready to be deployed. Of the 71 PIs elicited, 13 were applicable to all of the MECs, and 58 were specific to a single MEC.

We found that the Instructor Pilots were able to generate concrete, specific descriptions of behaviors that indicated low levels of readiness (e.g., does not have radar properly set up, uses extraneous communication). Interestingly, the instructors most easily defined high levels of readiness as being the absence of commonly occurring errors, rather than as the presence of defined expert behaviors. When asked to describe the behaviors that characterize a highly expert four-ship team, the experts replied, essentially, “The absence of the errors that we have just

described.” This may be specific to the air-to-air combat environment, which is demanding and requires a high degree of expertise to perform multiple tasks simultaneously at acceptable levels.

Find “Rich” Areas for Assessment

The overall goal of the SPOTLITE project was to focus performance measurement on those behaviors that are most diagnostic of overall team performance. As an initial top-down identification of important areas for measurement, we set up a matrix that arrayed the 71 PIs as rows and the knowledge and skills as columns, and asked five SMEs to rate the importance of each knowledge or skill for each behavior. The data from the rating matrix were then analyzed from two different perspectives. First, we identified those behaviors (PIs) that drew on many different types of knowledge and skills. We consider these behaviors to be “rich” areas for assessment because successful performance of the behavior indicates that the learner possesses many different types of knowledge and skills. Our goal was, to the extent possible, to include these behaviors in the SPOTLITE measurement instrument because we expected them to be diagnostic of the competency levels of the four-ship. We also analyzed the rating matrix from the viewpoint of identifying the knowledge and skill elements that were most important in this simulation environment with a similar goal of ensuring that these critical knowledge types and skills were, to the extent possible, assessed by the behaviors to be included on the instrument.

The analysis of rich areas for assessment and critical knowledge and skills, based on the MECs, provided us with a top-down perspective on what should be measured by the SPOTLITE instrument. The next step was to work from a bottom-up perspective to identify those behaviors that could be measured in the simulation environment.

Identify Behaviors Observable in the Simulation-Based Training Environment

The SMEs identified PIs based on their experience in many different environments, including actual flight as well as simulation-based training. In developing measures for the training facility at Mesa, we wanted to ensure that we did not include (a) behaviors that, although important, are not possible to observe in the simulation environment; (b) behaviors that are so routine that there is no variation among teams in their occurrence; or (c) behaviors that are so rare that they are hardly ever observed (i.e., floor and ceiling effects). Our goal was to identify behaviors associated with effective performance that are sometimes, but not always, observed in the training simulation and on which performance varies among the four-ship teams who train in the facility.

For this stage of measure development we relied heavily on the playback facilities of the training simulation facility. We selected a benchmark scenario that is used for almost all training sessions in the facility and accessed the recorded performance of multiple four-ship teams flying in this benchmark scenario. Using the playback capability, we reviewed the performance of multiple teams flying the same scenario, working closely with SMEs to identify the desirable and undesirable behaviors that occurred for each team and making detailed notes regarding the nature of the behavior and when it occurred in the scenario. At the end of a one-week session, we had accumulated a large body of material on the behaviors (both positive and negative) that were observable in the simulation, behaviors that had actually occurred in the training facility, and when in the benchmark scenario these behaviors had occurred. These observed behaviors provided the raw material for initial measure development.

Develop and Test Draft Measures

Based on the lists of behaviors that were both observable and observed in the simulation environment, we prepared diagrams that summarized the observed behaviors in the context of the benchmark scenario. The next step was to collapse these diagrams into behaviorally anchored rating scales associated with points in the scenario. The diagrams were helpful in capturing the language used by SMEs to describe observed behaviors, allowing us to produce a draft of the rating scales that used clear, meaningful language to describe the end points of the scales. Clear descriptions of the anchor points on the rating scales are essential if the instrument is to be reliable when used by multiple SMEs.

As noted previously, our planned approach in developing the instrument was to tie each item in the instrument to a specific point in the scenario. However, because of the dynamic nature of the scenarios, few invariant points can be specified a priori. Rather, the way in which a scenario unfolds and the events that happen are in large measure dependent upon the actions taken by the team. Thus we found that although we could not specify a particular point at which the measurement should be taken, we were able to designate a specific interval, or segment, in the scenario in which the measurement should be taken. For example, we could specify that an item should be assessed during the initial detection phase of a scenario before any enemy groups are targeted, during the targeting and engagement phase, or during the assessment phase of an engagement.

As we developed the wording, ordering, and content of the items for each segment of the scenario, we applied the instrument to multiple teams to test the process and determine real-time rating feasibility. We found that initial versions of the instrument were too complex and too long for feasible use by one observer during a training session. To shorten and simplify the

instrument, we eliminated a number of items that, although they could be observed, also could be obtained directly from the simulation environment, such as whether or not the shots taken were valid. The goal was to focus the expert observer's attention on those aspects of performance that could not be assessed easily from data obtained directly from the simulation, such as communication. Performance data that can be obtained directly from the simulation environment are captured in a parallel effort (Schreiber et al., 2003).

Although our goal in developing the SPOTLITE measures was to rate the team as a whole, not individual members, as we reviewed teams performing in the scenario we found that, in response to enemy maneuvers, the four-ship often split into two elements that could be involved in different types of engagements or even in different phases of a mission. In these cases, the SMEs reported that an integrated evaluation of the whole team was not meaningful assessment. Rather, they wanted to evaluate each element independently. They were, in effect, evaluating two simultaneous engagements, and we needed to design the instrument to accommodate this situation. The solution we suggested was that, depending upon the actions of the team, observers could rate the four-ship team as a whole, or they could rate the two elements separately.

Generalizing the Instrument to Multiple Scenarios

When we started developing the instrument, we decided to focus on one particular benchmark scenario, make the instrument appropriate for that scenario, and then consider whether it could be generalized to other scenarios. We structured the flow of questions around the events that were preprogrammed into the scenario—typically preprogrammed actions that the Red aircraft would take (e.g., breaking a formation, turning back towards the friendly forces, etc.). Therefore, in the original version of the instrument, the questions were linked to specific

enemy actions that appeared in the benchmark scenario we used for initial development. When we showed our initial prototype to SMEs, they suggested that we should organize the instrument around types of engagements rather than around specific enemy actions or groups. Organizing the form around types of engagements of friendly aircraft with enemy groups was a major advancement in that it facilitated development of an instrument that is appropriate for any scenario, rather than an instrument that must be tailored for each scenario. Equally important, it focused the evaluation on the actions of the “friendlies” (the actual pilots being evaluated) rather than the opponents, which are simply computer-generated forces.

We classified engagements into two types: Beyond Visual Range (BVR) and Visual Identification (VID). We found that a majority of our items applied to both BVR and VID engagements. In a few cases, the stem of the item applied, but the behavioral anchors were different for the two types of engagements. A few specific items applied to only one type of engagement. For example, whether the flight achieved proper fan/bracket geometry is appropriate for a VID engagement but not for a BVR engagement. At the end of the development phase, the instrument included 26 items that spanned the phases of the mission and the relevant areas of competence.

Software Implementation of the SPOTLITE Instrument

Our goal was to implement the SPOTLITE instrument on a handheld computer for ease of use and rapid entry of the assessment data into a computer database. A computer-based implementation was potentially much more flexible than a paper-based implementation, allowing the observer to quickly add sections and pages in real time as the engagement unfolds. To achieve this goal, we designed the user interface to support fast and accurate data entry by observers during a scenario.

Figure 3 shows a snapshot of the instrument as it was implemented with the split screen that allows an observer to rate two engagements simultaneously. In the example, on the left side of the screen the observer is in the process of answering questions about the targeting/engagement phase of a BVR engagement involving the 1-2 element. In the second question (regarding targeting), the observer has clicked on 2, indicating that Pilot 2's performance was weak. The red pencil icon indicates that the observer has made a note about that question. On the right side, the observer is responding to questions pertaining to the assessment phase of a VID engagement involving the 3-4 element. The "New BVR" and "New VID" buttons are active on the right side because the observer can bring up a new engagement after completing an assessment page. They are not active on the left side of the screen because the observer has not yet filled out the assessment page for that engagement, required before bringing up a new engagement screen.

The screenshot displays the Observer Questionnaire v1.1 interface, split into two main sections: BVR 1 Targeting and Engagement 1/2 and VID 1 Assessment 3/4. Each section contains several questions with progress bars and response indicators.

BVR 1 Targeting and Engagement 1/2:

- RADAR:** How appropriately are the RADARs set up? Progress bar is at 4. Response indicators: 1, 2, 3, 4.
- TGT:** Does the flight target in accordance with standards? Progress bar is at 3. Response indicators: 1, 2, 3, 4. A blue arrow points to this question with a callout box labeled "Red Pencil".
- BDT RX:** Do the fighters recognize and react to bandit maneuver? Progress bar is at 3. Response indicators: 1, 2, 3, 4.
- SHOTS:** Does the flight shoot in accordance with shot doctrine? Progress bar is at 4. Response indicators: 1, 2, 3, 4.

VID 1 Assessment 3/4:

- FORM:** As the team comes out of the engagement, to what extent do they have mutual support? Progress bar is at 4. Response indicators: 1, 2, 3, 4.
- COLD OPS:** Posit, picture, plan during cold ops: Progress bar is at 5. Response indicators: 1, 2, 3, 4.

The interface includes buttons for "Status", "Abort", "Picture Clean", "New BVR", "New VID", "Assess", "Notes", and "Cancel".

Figure 3. Example of BVR target/engagement page and VID assessment page with responses, progress bars, and notes icon.

When the observer has completed the last page of the instrument, a status page appears and shows pages in which some of the questions have not been answered. At this point, the observer can go back to any of the incomplete pages and fill in unanswered questions—but the observer is not required to complete all questions. When the observer has responded to all or as many of the questions as desired, he or she clicks the “submit” button, and the observer’s responses are stored to a file.

Final Measure Evaluation

Criteria for Evaluation

For the SPOTLITE Instrument to be useful for the purposes of evaluation of team performance, it had to meet a number of criteria: proper behavior identification and measurement, sensitivity, reliability, and validity.

First, the instrument must tap the behaviors that are relevant for performance in the domain. We met this criterion by identifying the PIS and working closely with domain SMEs to specify relevant, observable behaviors. We then linked the items that were developed to the underlying knowledge and skills that have been identified as critical for high-quality performance in the air-to-air combat domain.

Beyond domain relevance, to be useful in differentiating performance, the instrument should be sensitive to varying levels of performance. If all the teams receive the same rating on an item, it is not useful for differentiating performance. As well as ensuring variations in scores, we wanted to insure that we did not include items for which there was a ceiling effect (i.e., an item on which most teams score very close to the top) or a floor effect (i.e., an item on which most teams score very low).

In addition to ensuring that the measures tap relevant factors and provide a range of scores, the data from the instrument needed to be reliable. In this context, we are concerned with inter-rater reliability, the extent to which different observers give a team consistent ratings on the items. To make comparisons across teams or within a team over time, when different raters may be assessing the team's performance, any variability in ratings must be due to performance rather than the difference in raters. We used standard psychometric techniques for assessing reliability (Nunnally, 1978).

A measurement instrument may be reliable in that raters concur on ratings, but still may not provide a valid measure of performance. For this application, validity has to do with whether the results from the instrument are consistent with another accepted measure of performance quality. The challenge in this regard was to identify an acceptable criterion variable. To address this challenge, we used an approach developed by Serfaty, MacMillan, Entin, and Entin (1997) for assessing the validity of measures of tactical performance: The criterion measure was an overall rating by SMEs of the individual's tactical expertise. For the SPOTLITE instrument, the criterion measure was the SMEs' overall ratings of the four-ship team's performance on the scenario on a 5-point scale that ranged from "novice" to "expert."

Experiment Design for Measure Evaluation

We conducted an experiment to assess the SPOTLITE instrument in terms of the evaluation criteria described above. We selected recorded data for the performance of 10 teams on the benchmark scenario that was used during the development of the SPOTLITE instrument. In selecting the teams, we tried to obtain range in performance by including highly experienced teams as well as less experienced teams, but the experience levels of the teams were not known to the SMEs who served as raters in the experiment.

Using the replay facility available at the Mesa lab, six SME observers independently rated each of the 10 teams using two approaches: (a) in real time using the SPOTLITE instrument and (b) at the end of the scenario based on overall team performance on the mission. Each observer rated 8 of the 10 teams using both the SPOTLITE instrument and the overall measure of performance. They rated 2 of the 10 teams on the overall measure only. The reason why they did not rate all of the teams using the SPOTLITE instrument is that we needed to

ascertain whether the SMEs' ratings on the overall measure were systematically influenced by their having rated the individual items in the SPOTLITE instrument.

Results

To conduct the assessment we computed an average score on the SPOTLITE instrument, the mean of the 26 items in the instrument. We refer to this as the *instrument mean*.

Impact of SPOTLITE instrument on overall measure of performance. Each observer rated 8 teams using both the SPOTLITE instrument and the overall rating and rated 2 teams using just the overall rating. The assignment of the teams that each observer rated overall was randomly determined with the constraint that the overall only ratings encompassed all 10 teams.

To ascertain whether we had gathered evidence that completing the SPOTLITE instrument systematically biased ratings on the overall measure, we compared the mean ratings based on the observer's overall ratings to the mean based on cases in which the observer completed both the SPOTLITE instrument and the overall rating. The means and standard deviation for the two sets of ratings are shown in Table 2. We did not find any significant differences between the mean ratings for the two conditions ($t = 0.62$, $df = 58$, ns). Hence we concluded that completing the SPOTLITE instrument items did not systematically bias the overall rating upward or downward, and therefore in the analyses that follow we did not distinguish between the overall ratings done without the SPOTLITE instrument and those done along with the SPOTLITE Instrument.

Table 2

*Comparison of Mean Overall Scores in Conjunction With and Independently of SPOTLITE**Instrument*

	<i>n</i>	Mean	Standard deviation
Ratings with SPOTLITE	48	2.95	0.95
Ratings without SPOTLITE	12	2.75	1.18

Note. $T = 0.62$, $df = 58$, ns

Sensitivity of measures. The means and standard deviations for the 10 teams on the SPOTLITE Instrument and the mean overall ratings of the 10 teams are shown in Table 3. As Table 4 shows, the mean scores on the instrument ranged from 2.43 to 4.09; the overall measure ranged from 1.50 to 4.25. The range of means indicated that both measures were able to differentiate the teams, and that the measures showed neither a floor nor a ceiling effect. The standard deviation across the teams (0.55 for the mean score and 0.91 for the overall score) was higher than the mean standard deviation across the raters for the individual teams, which also supported the sensitivity of the test to differences in team performance.

Table 3

Means and Standard Deviations for 10 Teams on SPOTLITE Instrument and Overall Rating

Team	Number of observers	Instrument mean	Instrument Std. Deviation	Number of observers	Overall	
					rating mean	Overall rating Std. Deviation
1	5	3.63	0.12	6	3.25	0.76
2	5	4.09	0.31	6	3.67	0.41
3	4	4.07	0.30	6	4.25	0.52
4	5	3.07	0.19	6	1.75	0.61
5	5	3.90	0.23	6	3.41	0.20
6	4	2.43	0.20	6	1.50	0.45
7	4	3.80	0.27	6	2.83	0.26
8	4	3.11	0.45	6	2.50	0.77
9	6	3.91	0.33	6	3.75	0.27
10	6	3.14	0.25	6	2.17	0.41
Mean	48	3.53	0.56	60	2.91	0.99

Reliability. To assess inter-rater reliability we computed a coefficient alpha for each team. Alpha levels above 0.8 are considered acceptable (Nunnally, 1978). The coefficient alpha for the SPOTLITE instrument based on six observers ratings of 10 teams was 0.98. The coefficient alpha for the overall performance measure was .95. These data showed that we were successful in developing behaviorally anchored scores with a high degree of reliability. This

means that the performance ratings for teams were not dependent upon the particular observer making the judgments.

To examine reliability further, we computed two categories of subscores, one by phase of mission and one by type of action, and assessed these subcategories to insure that reliability was maintained. As an example, we computed reliability ratings for BVR and VID engagements separately and found similarly high reliability scores (.98 and .97, respectively).

We also analyzed the data to determine whether each item met sensitivity and reliability criteria. Figure 4 shows the mean ratings and the range in ratings for four teams on one of the SPOTLITE items (mutual support). The boxes in the figure show the mean scores assigned by the six observers for 4 teams. As shown in Figure 4, the scores for the teams ranged from 1.0 (novice) to 4.8 (expert). On the other hand, the ranges of scores within the teams, shown by the arrows, were much lower. In the case of the low-performing team, all six observers assigned a rating of 1. In the case of the highest performing team, the observers' scores ranged from 4.5 to 5. The scores for the next highest performing team ranged from 4 to 5. For the team that performed just below the midpoint, the scores ranged around the midpoint of the scale, from 2 to 4. The within-team mean range was 1.2. Thus, there was much lower variability in the team scores across the raters than in the range of scores across the teams. Figure 10 also shows that the full range of score values was used, suggesting that this item is sensitive to differences in performance across the teams, and that the raters were reasonably consistent in their assessment of each team's performance.

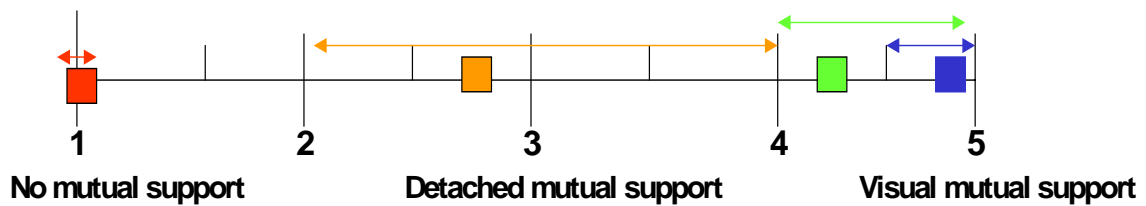


Figure 4. Range of ratings for 4 teams on the item: To what extent does the team have mutual support?

Validity. Because there is no “gold standard” or external performance criterion against which to measure the SPOTLITE-determined ratings, we used a measure of construct validity. To assess the validity of the scores on the SPOTLITE instrument, we compared them to the overall ratings given by the SMEs. Figure 5 shows the relationship between the overall ratings for the teams and the mean score on the SPOTLITE instrument assigned by each SME for each team. The correlation between the two measures (and the standardized beta coefficient) was .79 ($n = 48, p < .0001$). The correspondence indicated that taken as a whole the SPOTLITE instrument items provided a valid measure of overall team performance as well as reliable measures of performance on specific aspects of behavior.

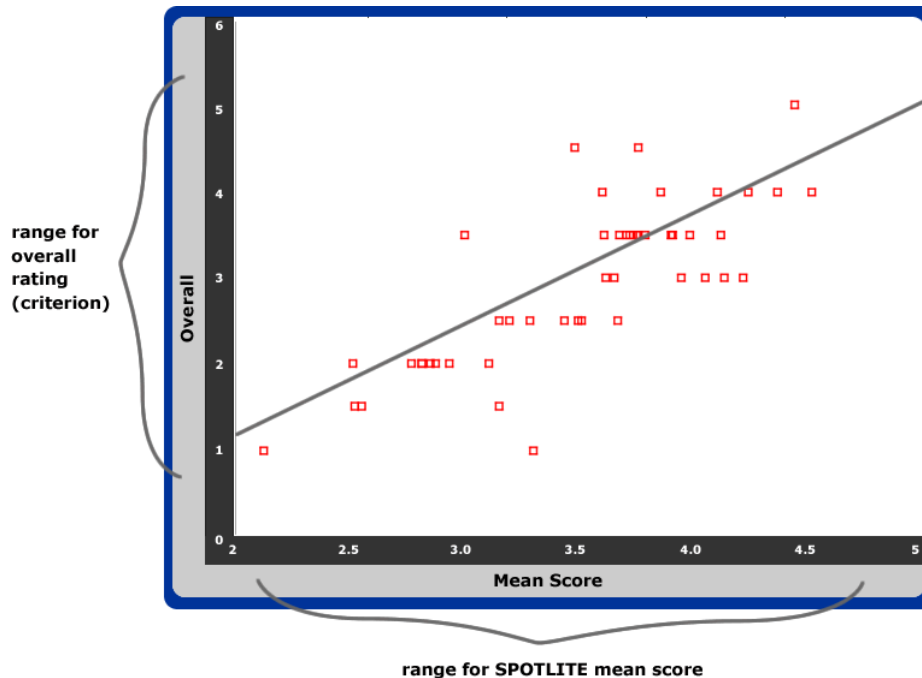


Figure 5. Relationship between overall rating and mean score on SPOTLITE instrument.

Discussion

Convergence on SPOTLITE Items

Our work confirmed that development of a reliable and valid assessment instrument is a labor-intensive process involving multiple iterations, and not just a linear process of developing, pilot testing, and refining questions. We studied missions performed by a variety of teams to insure that the instrument covered dense, diagnostic points in the scenario and tapped critical knowledge and skill elements. As we developed the instrument, the review of each new team stimulated new questions or potential modifications to existing ones. As we considered new questions, we had to insure that a new candidate question was not redundant, was applicable across scenarios, was diagnostic of important knowledge and skills, and tapped elements of performance that could not be captured by the system. We found that because the process extended over a period of time, questions reemerged that were previously included and then rejected.

To insure that we have covered the range of MECs and the critical knowledge and skill elements identified in our initial analyses, we constructed a matrix of instrument items by knowledge and skill elements (analogous to the PI by knowledge and skills) and asked the SMEs to indicate which knowledge and skill elements each item assessed. We found that the questions covered all the MECs to the extent that they can be assessed in a simulated environment (e.g., force reconstitution is not well tested in the simulation) and that the critical knowledge and skill elements were addressed by one or more of the questions.

A number of factors contributed to the success of this effort and would be important for analogous efforts. One of the most important is the full cooperation of the highly expert SMEs who supported us in developing the SPOTLITE instrument. Five SMEs devoted considerable effort to this project both in identifying the best measurement points and in helping us to develop meaningful questions and diagnostic behavioral anchors.

Issues in Real-Time Assessment

Our objective in developing the assessment instrument was to have observers provide their ratings in real time, based on specific aspects of the team's performance in a specific, narrow segment of the scenario. Even with a pared-down set of questions, however, completing the instrument in real time is extremely challenging for observers. Significant events happen in seconds, or even fractions of seconds, because of the fast-paced action. Observers must pay attention to the four members of the team, each of whom has four different cockpit displays. When an observer is responding to an item, he or she could miss team members' actions and their communication. We found that this is most likely to happen when the team has split into two elements, the two elements are in different types of engagements, and the observer needs to watch both elements simultaneously. For example, one element may be involved in a BVR

engagement and the other in a VID engagement, or one element may be engaged while the other is in an assessment phase. For this reason, we established the convention that observers should omit any items that they could not rate because they did not see the relevant actions. However, in the process of testing the instrument, we found that the SME observers were unwilling to use a “did not see” option. Hence, SMEs suggested being able to fill in the item during the debriefing, when the recorded mission was replayed so that the team could analyze and discuss their performance. Observers felt that their assessment would be more accurate given one of two additional opportunities: (a) bringing back the instrument during the debriefing, when the team was reviewing the scenario and the observers could rate behaviors they had missed in real time, or (b) having two layers of questions, one completed in real time and the other during the debriefing or during a controlled replay of the scenario. These alternatives are being explored for future versions of the assessment instrument.

Observer Training

Use of the instrument for evaluation over the course of several months revealed the need for careful attention to training so that all observers who use the instrument fully understand the scales, including the rating points that do not have specific behavioral anchors. We found this especially important for those scales that involve two concepts, such as appropriateness and effectiveness of use of a particular technique. Use of the instrument over time has shown that observers need training in how to apply the instrument based on the consensus developed by SME observers. For example, new observers need to be clear about what constitutes a “new” engagement. Use of the instrument also has revealed the need for training in what might be called nonstandard situations. For example, observers need training in how to handle a mixed

BVR–VID engagement. Other conventions that need to be addressed in the training manual include how to handle situations in which fighters are shot down by the enemy.

Conclusions

In this project we demonstrated the feasibility of developing a rating instrument that allows observers to make valid, reliable ratings of performance in a complex, dynamic, multiperson task and to implement that instrument for real-time use on a handheld computer. The SPOTLITE instrument we developed yields quantitative data useful for comparison across teams and across time within a team. Grouping items by phase of the mission or by type of action (e.g., maneuver, formation, etc.) allows for a more precise diagnosis of team performance. The use of positive or negative “checkmarks” for individual members of the team whose performance on a particular item is significantly above or below the rest of the team provides a mechanism for observers to denote variance in level of competency across the team. Scoring of individual pilots as excelling or deficient can be used in feedback and also in analysis to determine the relationship between homogeneity of teams and performance.

The approach we used to develop the SPOTLITE instrument is applicable to other domains, but requires SME support for identifying the most diagnostic measurement points and for developing the questions, including the rating items and the behavioral anchors. Once an initial draft has been developed and agreed upon by multiple SMEs, the instrument still requires real-time testing to insure that it is usable and applicable across a range of missions and levels of performance. With future research and development of similar instruments, observers can rate complex multiperson tasks more accurately to contribute to better performance.

References

- Baker D. P., & Salas E. (1997). Principles for measuring teamwork: A summary and look toward the future. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 331-355). Mahwah, NJ: Lawrence Erlbaum.
- Brannick, M. T., Salas, E., & Prince, C. (Eds.) (1997). *Team measurement and performance: Theory, methods, and applications*. Mahway, NJ: Lawrence Erlbaum.
- Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Volpe, E. (1995). Defining team competencies and establishing team training requirements. In R. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations*. San Francisco: Jossey Bass.
- Colegrove, C. M., & Alliger, G. M. (2002). *Mission essential competencies: Defining combat readiness in a novel way*. Paper presented at the SAS-038 NATO working group meeting, Brussels, Belgium.
- Dwyer, D. J., Fowlkes, J. E., Oser, R. L., Salas, E., & Prince, C. (1997). Team performance measurement in distributed environments: The TARGETS methodology. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp 137-154). Mahwah, NJ: Lawrence Erlbaum.
- Entin, E. B., & Entin, E. E. (2000). Assessing team situation awareness in simulated military missions. In *Proceedings of the 44th annual meeting of the Human Factors and Ergonomics Society, San Diego, CA* (pp. 73-76). Santa Monica, CA: Human Factors and Ergonomics Society.
- Johnston, J. A., Smith-Jentsch, K. A., & Cannon-Bowers, J. A. (1997). Performance measurement tools for enhancing team decision making. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Team*

- performance assessment and measurement: Theory, methods, and applications* (pp. 311-327). Mahwah, NJ: Lawrence Erlbaum.
- MacMillan, J., Paley, M., Entin, E. B., & Entin, E. E. (2004). Questionnaires for distributed assessment of mutual team awareness. In *Handbook of human factors and ergonomics methods*. Winter Park, FL: Taylor and Francis.
- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw Hill.
- Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, 52, 471-499.
- Schreiber, B. T., Watz, E. A. & Bennett, W., Jr. (2003). Objective human performance measurement in a distributed environment: Tomorrow's needs. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*. Orlando, FL.
- Serfaty, D., MacMillan, J., Entin, E. B., & Entin, E.E. (1997). The decision-making expertise of battle commanders. In C. E. Zsanbok & G. Klein (Eds), *Naturalistic decision making*. Mahwah, NJ: Lawrence Erlbaum.
- Smith-Jentsch, K.A., Johnston, J. H., and Payne, S. C. (1998). Measuring Team-Related Expertise in Complex Environments. In J. A. Cannon-Bowers & E. Salas (Eds), *Making decisions under stress*. Washington, D.C.: American Psychological Association.
- Smith-Jentsch, K., Payne, S.C., and Johnston, J. H. (1996). Guided team self-correction: A methodology for enhancing experiential team training. In K. A. Smith-Jentsch (Chair), *When, how and why does practice make perfect?* Paper presented at the 11th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Notes

1. The views reflected in this paper are those of the authors and should not be construed as representing the official position of their respective organizations.

2. The work described in this paper was conducted under contracts No. F41624-99-C-6025 and F33615-00-C-6004 and with Air Force Research Laboratory, Mesa, Arizona.