

A Human Systems Integration Method for Validating Team Performance Assessment Within a Simulation-Based Training System

Joan H. Johnston, Dennis A. Vincenzi, & Paul H. Radtke
NAVAIR Orlando Training Systems Division, Orlando, FL

William Salter & Jared Freeman
Aptima, Inc., Woburn, MA

Improving simulation-based training is the focus of a Navy program of research titled Debriefing Distributed Scenario-Based Exercises (DDSBE). DDSBE is intended to optimize training effectiveness by demonstrating improved methods and technologies for assessing team and multi-team performance, diagnosis routines, and debriefing templates. A main objective of the DDSBE program in spiral 1 was to test the training simulator capability to produce valid and reliable team performance assessments. This paper describes a human systems integration test methodology and an exploratory experiment that applied the methodology. The results showed excellent assessment validity and inter-rater reliability when a hand-held PC was used to assess team task and teamwork behaviors in conjunction with an automated performance recording capability. The results of this analysis should provide guidance to simulation developers in designing distributed training that supports learning combat team skills.

Background

The Navy plans to greatly increase opportunities for advanced teams to train and practice team-level mission objectives by connecting multiple simulation systems across distributed sites. To achieve this vision, new training systems must have the capacity to accurately assess sailors' acquisition of both individual and team-based knowledge and skills. In many cases, however, fielded simulations lack embedded human performance assessment tools to assess performance at any level. Without technical support from the training system, performance assessment, diagnosis, and debriefing become the responsibility of instructors and the other personnel supporting the Navy's training programs. This burden on the instructor is made more difficult during distributed mission training exercises, in which several teams within different platforms train together in a combination of live, constructive, and simulation-based systems (Neville, Fowlkes, Milham, Merket, Bergondy, Walwanis, & Strini, 2001). These new requirements greatly increase the burden to address multi-team performance in non-face-to-face interactions between instructors and training teams.

Improving the embedded assessment capabilities of simulation-based training is the focus of a Navy program of research titled Debriefing Distributed Scenario-Based Exercises (DDSBE; Johnston, Radtke, Van Duyne, Stretton, Freeman, & Bilazarian, 2004). The four-year project is led by NAVAIR Orlando Training Systems Division and is funded by the Office of Naval Research. The DDSBE program will deliver training technology solutions for use in the Navy's plan to conduct distributed simulation-based training involving naval aviation post Fleet Replacement Squadron teams.

A major technical objective of the DDSBE program is to design a testbed that can be used to test the effectiveness of debriefs and after action review interface methods and designs for single and multi-combat teams. This system must produce

valid and reliable assessments of team performance so that diagnosis and after action review can be effective. Therefore, a test methodology was developed to determine how valid and reliable the testbed's data collection and assessment systems are before further development of the diagnosis and After Action Review (AAR) capabilities. This approach was intended to support the DDSBE software development team in building mechanisms for the right kinds of team performance observations and for the accurate integration and transformation of observations into measurement. Findings from this exploratory experiment would then enable future experimentation to compare current approaches to training with an improved capability.

Approach

The experimental design was a 3x3 factorial and used within-subjects comparisons to observe the sensitivity of the DDSBE measurement system when trainees perform at different levels of acceptability. The testbed included two E-2C aircraft mission stations, linked through a local area network to form a distributed simulation training system. The network was operated by roleplayers representing two E-2C Naval Flight Officers (NFO); the Combat Information Commander (CICO) and Air Combat Officer (ACO); and a third roleplayer who performed as the F/A-18 Strike Leader and other external communication sources. An unclassified 40-minute strike mission scenario was developed based on selected Navy Mission Essential Tasks and related team training objectives. Subject Matter Experts (SMEs) vetted the scenario, identified critical events, and described the expected behaviors of the two E-2C NFOs working at each performance level in the scenario to be used in the assessment.

Table 1 presents the factorial design. For the nine scenarios the two NFO roleplayers' performance levels were manipulated (i.e., less than acceptable, minimally acceptable, above acceptable) on selected taskwork and teamwork

behaviors to judge DDSBE system sensitivity. For example, during Scenario 5, both role-players performed at the “acceptable” level, whereas in Scenario 6 the ACO performed at the “above acceptable” level and the CICO performed at the “acceptable” level. The Scenarios were presented in random order, as indicated in the table.

Table 1. Performance Level of ACO & CICO by Scenario Number and Presentation Order.

	ACO Less Than Acceptable	ACO Acceptable	ACO Above Acceptable
CICO Less Than Acceptable	Scenario 1 Order 6	Scenario 2 Order 9	Scenario 3 Order 8
CICO Acceptable	Scenario 4 Order 2	Scenario 5 Order 5	Scenario 6 Order 7
CICO Above Acceptable	Scenario 7 Order 1	Scenario 8 Order 4	Scenario 9 Order 3

Previous research suggested that accurate assessment of team and task performance could be based on a sample of verbal communications and their associated mission station interactions (Smith-Jentsch, Johnston, & Payne, 1998). The nine 40-minute scenarios were designed to elicit specific team verbal communications and mission station behaviors. In all, 38 specific trigger events were inserted into the scenarios, including 14 events that were designated as “key” or “critical.” Behaviors for each scenario were then scripted for the roleplayers to manipulate their level of performance. The nine scenarios each contained between 193 and 223 possible performance scoring opportunities.

The two E-2C mission stations were instrumented to automatically capture and record operator keyboard and mouse inputs. A hand-held PC tablet computer, the Virtual Communications Assessment Tool (VCAT), was used by each member of the SME assessment teams to record verbal communications (Stretton & Wilson, 2004). Figure 1 presents the interface used by the VCAT-aided evaluators.

Two teams of two persons each were created to permit evaluation of the systems interrater reliability.

The two-person VCAT-aided assessment teams divided responsibilities for observing and recording specific types of performance. One evaluator made alert-based notations of team members’ communication accuracy and timeliness for each critical event, and made assessments of the team’s “leadership and initiative” and “supporting behaviors.” The second evaluator assessed accuracy and timeliness of communications between the E-2C NFO team and the F/A-18 pilot roleplayer, as well as two teamwork behaviors of “communication” and “information exchange.” (See Smith-Jentsch, Zeisig, Acton, & McPherson, 1998.)

The DDSBE system sent the VCAT visual alerts when one of the pre-selected scenario events was scheduled to

occur. The alerts prompted the evaluator to listen to the tactical communications and evaluate them through the VCAT interface. Performance on teamwork behaviors (i.e., leadership/initiative, supporting behavior, communications and information exchange) was not tied specific events and could be observed and rated at any time. As a manipulation check, a fifth SME evaluator provided an independent assessment of the team’s tactical and team performance for each of the 38 trigger events using a paper-based tool.

The DDSBE system integrated the automatically recorded team member actions and the human-entered VCAT assessments via the Automated Performance Assessment (APA) capability (Carolan & Bilazarian, 2004) that produced a report that formed the basis for determining whether team members achieved the pre-specified training objectives in the scenario. The six training objectives were:

- Conduct anti-air combat detection and identification
- Execute correct target area tactics
- Conduct off count
- Make strike reports to higher authority

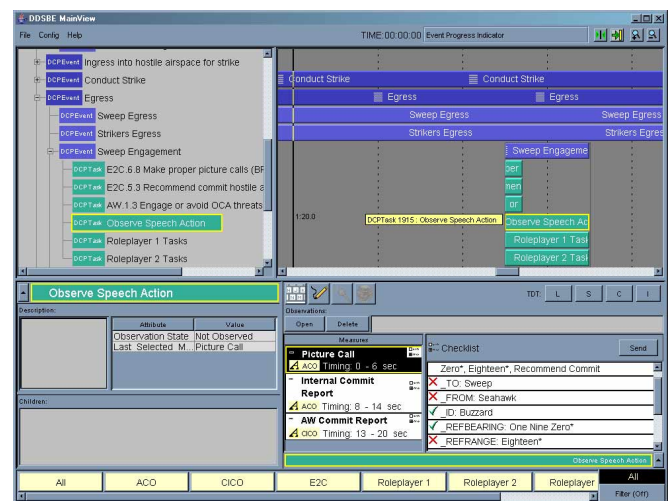


Figure 1. VCAT Interface

- Ensure strike mission proceeds according to the planned timeline
 - Perform surface-to-air (SAM) threat reporting
- Separate reports were produced for the two VCAT-aided two-person assessment teams.

Procedure

The experiment was conducted over two days in October 2004. On Day 1, the three roleplayers and five evaluators completed informed consent forms and were trained on the kinds of measures they were expected to take, how to use the measurement tools, the mission being represented, and the events in the scenario they were to observe. The introductory briefings were followed by two familiarization practice exercises to train the evaluators to use the VCAT. The familiarization sessions used selected portions of the experimental scenario.

Four scenario runs followed the familiarization session. Evaluators were blind to the experimental manipulations throughout the experiment. The evaluators observed the two roleplayers performing the scenario and recorded their observations and ratings. After each scenario, evaluators also completed a TLX workload self-assessment questionnaire (Hart & Staveland, 1988). Day 2 followed the same schedule for the remaining five data collection runs. Following the final run, the evaluators were asked to rate the usability of VCAT and their reactions to the DDSBE system during a structured group interview. All participants discussed the process, the measures, and the workload of being an evaluator. The discussions also gathered suggestions for improvements and applicability of the DDSBE system to other Navy team-training and research efforts.

Results

The analysis focused on three questions:

- Were the assessments of the team’s performance that was produced by the DDSBE system valid?
- Were the assessments of the team’s performance reliable across the two VCAT-aided assessment teams, and in relationship to the independent, unaided assessor?
- Were there any patterns or differences in the workload experienced by the evaluators?

Validity of Taskwork Scores. The analysis of the DDSBE system’s validity and reliability focused on the 14 “key” and “critical” events in the scenario. The team’s aggregate performance on these fourteen events determined how well the team accomplished the six designated training objectives listed earlier.

We compared the integrated performance scores computed by the DDSBE system for each event based on the APA and VCAT inputs with the scores that should have been awarded based on the actual scripted performance of the roleplayers. A high correlation was interpreted to mean that the performance scores recorded by the assessment teams were valid indicators of the actual performance they were observing. Overall, correlation coefficients were computed for 117 valid observations for (14 events across nine scenarios, minus nine missed observations). Table 2 presents the correlation coefficients between the scripted scores and the scores produced by the two VCAT –aided teams, and by the independent overall evaluator. The observations for all three sets of assessments correlated significantly with the scripted performance. However, the observations of the two VCAT-aided Assessment teams were significantly more reliable than those of the un-aided overall evaluator.

Table 2. Correlations Between Observed and Scripted Scores on “Key” and “Critical” Scenario Events

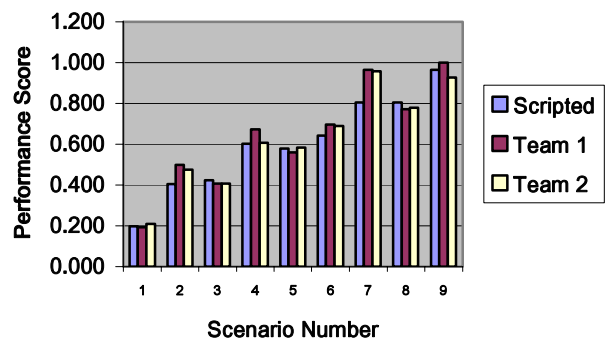
	Correlation Coefficients with “Scripted” Scores
VCAT Team 1 Scores	$r = .801$ (sig. < .005, df=117)
VCAT Team 2 Scores	$r = .799$ (sig. < .005, df=117)

Overall Evaluator Scores	$r = .244$ (sig. < .005, df=117)
---------------------------------	-------------------------------------

Similar results were achieved with respect to the assessment of the six training objectives. For example, Figure 2 presents the aggregate performance scores for one of the six training objectives, *conduct anti-air combat detection and identification*, that were produced by the two VCAT-aided assessment teams compared to the performance scores based on the team’s actual performance. In this instance, the performance scores across the nine scenarios improved more or less monotonically with the scripted performance.

The score for this training objective is an aggregate of performance observed over four of the fourteen key and critical events. Figure 2 illustrates the similarity between the scripted and the recorded performance for both VCAT-aided assessment teams. The performance scores produced by VCAT Team 1 and 2 correlated .971 and .970 with the scripted scores on this training objective. By contrast, the scores produced by the independent evaluator correlated .525 with the scripted scores on this training objective¹.

Figure 2. Scripted and Recorded Scores for the Training Objective: Conduct Anti-Air Combat Detection and Identification



The training objective discussed above was essentially a team task because both members of the team could perform the steps required to meet the performance objective. However, some tasks, such the task of *making strike reports to higher authority*, were performed primarily by one of the team members, in this case, the CICO. Consequently, the scripted performance varied according to how well the individual team member performed rather than the team as a whole.

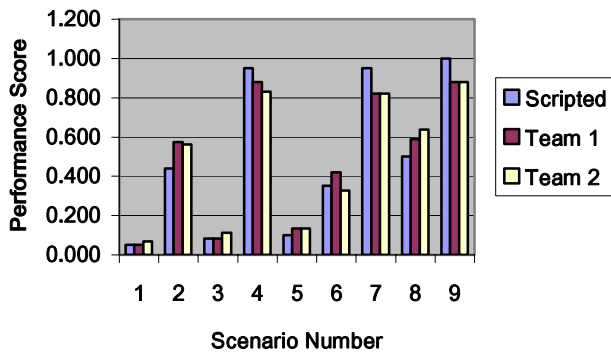
Figure 3 illustrates the ability of the DDSBE system to track performance on this type of task. The scripted performance score moves in a saw tooth pattern reflecting the change in the performance of the CICO from below acceptable, acceptable, and above acceptable.

¹ A direct graphical comparison of the VCAT scores and the scores produced by the independent evaluator could not be made because they were scored with different numeric scales.

The score for this training objective is an aggregate of performance observed over five of the fourteen key and critical events. The figure illustrates the similarity between the scripted and the recorded performance for both VCAT-aided assessment teams. The performance scores produced by VCAT Team 1 and 2 correlated .977 and .992 with the scripted scores on this training objective. By contrast, the scores produced by the independent evaluator correlated .392 with the scripted scores on this training objective².

Validity of Teamwork Scores. The analysis of the teamwork related scores focused on the entire scenario, since teamwork behavior could be performed at any time and may or may not be tied to a specific event. In addition, unlike the taskwork scores, the assessment teams were not prompted to make an observation relating to teamwork, but were free to record an observation whenever they felt that the performance was worth noting. Nevertheless, the validity of the teamwork scores awarded by the two VCAT-aided assessment teams was very high. There were between 69 and 85 potential teamwork related behaviors scripted into each scenario, of which the assessment teams recorded only a fraction. However, the overall assessment scores assigned by VCAT Team 1 for each of the nine scenarios correlated .867 with the actual percentage of possible behaviors performed. Similarly, the overall teamwork assessment scores assigned by VCAT Team 2 for each of the nine scenarios correlated .904 with the actual percentage of possible behaviors performed.

Figure 3. Scripted and Recorded Scores for the Training Objective: Making Strike Reports to Higher Authority



Reliability. Table 3 presents the intercorrelation among the scores produced by three sets of assessment teams for the 14 “Key” and “Critical” scenario events. The table also shows the correlation between the two VCAT teams with regard to overall teamwork over the nine scenarios.

The two VCAT-aided teams scored the fourteen events very similarly, as might be expected given the high correlation

of both teams with the scripted scores on these events, indicating a high level of both validity and reliability. Similarly, the overall teamwork scores were very similar, suggesting that the scoring methodology was reliable. However, neither VCAT team scored these events similarly to the independent evaluator.

Table 3. Correlations Between Scorers on “Key” and “Critical” Scenario Events: Taskwork and Teamwork

Comparison Scores	Correlation Coefficient
VCAT Team 1 X VCAT Team 2 (Taskwork)	.824
VCAT Team 1X VCAT Team 2 (Teamwork)	.926
VCAT Team 1 X Overall Evaluator	.297
VCAT Team 2 X Overall Evaluator	.302

Workload. Workload data were collected after each of the nine scenario runs. There was no apparent pattern of workload differences across scenarios, or between the two VCAT assessment teams. However, all workload subscales were rated higher by the member of the team who was monitoring the external communications scenario with the exception of temporal demand (See Table 4).

Table 4. Workload Reported By VCAT Team Assignment: Internal Vs. External Communications

	Internal	External	Diff
Effort	20.83	37.50	16.67
Performance	11.39	25.83	14.44
Frustration	8.61	30.00	21.39
Temporal Demand	33.33	35.00	1.67
Mental Demand	23.33	49.17	25.84
Physical Demand	8.06	13.33	5.28
Overall Workload	22.37	36.70	14.33

Internal communications raters experienced low to moderate workload while external communications raters experienced moderate to medium levels of workload. Internal communications raters experienced low levels of overall workload while external communication raters experienced moderate levels of overall workload. Although all differences were statistically significant except the difference between internal and external communications with respect to temporal demand, the greatest numerical differences appears to be experienced in the areas of frustration (diff = 21.39) and mental demand (diff = 25.84), indicating that the task of rating external communications was much more frustrating

² A direct graphical comparison of the VCAT scores and the scores produced by the independent evaluator could not be made because they were scored with different numeric scales.

and mentally demanding than the task of rating internal communications.

Discussion

This paper describes a relatively unique human systems integration methodology designed to assess the reliability and validity of a training simulator's human performance measurement capabilities before employing it as a training intervention. The initial results of this exploratory experiment indicate that the approach to collecting and combining team performance assessments via a hand-held PC resulted in excellent performance measurement validity and high inter-rater reliability. The observations of the VCAT-aided assessment teams also were significantly more accurate than those of the un-aided independent evaluator. The teamwork-related reliability was also high.

Equally important for this evaluation of the testbed was the fact that the system produced results that make sense from the standpoint of an instructor presenting a debrief to a team. The risk of reducing performance to an abstract score is that the number itself does not provide information about what the team or the individual member did to earn, or what would be required to improve the score. The value of the numeric score is that it alerts the instructor to significant areas of performance that require improvement, and provides the starting point for a deeper examination of the actual behaviors that produced the score. By structuring the scenario and the scoring system into specific events, the DDSBE system provides the basis for this "drill down" capability. Fortunately, this experiment indicates that the validity of the scores computed from the raw observational data is sustained at the level of the scenario, the training objective, and the individual event. The sampling approach used to measure the teamwork dimensions also appears to produce valid and reliable measures.

The next step in our analysis will be to assess the validity of combining the handheld recordings with more sophisticated automated data capture and analysis tool, the addition of the third NFO team member and a second external team – the F/A-18 Sweep Leader and wingman. The results of this additional analysis should provide guidance to simulation developers in designing distributed training that supports learning combat team skills.

References

- Carolan, T.F., & Bilazarian, P. (2004). Automated data collection and assessment for debriefing distributed simulation based exercises. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting [CD-ROM, p. 2567-2571]. Santa Monica, CA.
- Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P.A. Hancock and N. Meshkati, (Eds.), Human Mental Workload, (pp. 139 – 183), Amsterdam: North Holland.
- Johnston, J.H., Radtke, P. H., Van Duyne, L., Stretton, M., Freeman, J., & Bilazarian, P. (2004) Team training in distributed simulation based exercises. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting [CD-ROM, p. 2557-2561]. Santa Monica, CA.
- 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 [CD-ROM, p. 1035-1044]. NDIA, Arlington, VA.
- Smith-Jentsch, K.A., Johnston, J.H., & Payne, S.C. (1998). Measuring team-related expertise in complex environments. In J.A. Cannon-Bowers & E. Salas (Eds.), Making Decisions Under Stress: Implications for Individual and Team Training (pp. 61-87). Washington, DC: American Psychological Association.
- Smith-Jentsch, K.A., Zeisig, R.L., Acton, B., & McPherson, J.A. (1998). Team Dimensional Training: A strategy for guided team self-correction. In J.A. Cannon-Bowers & E. Salas (Eds.), Making Decisions Under Stress: Implications for Individual and Team Training (pp. 271-297). Washington, DC: American Psychological Association.
- Stretton, M.L., & Wilson, M.S. (2004). Semi-automated observation and assessment--trainer interaction within a distributed training environment. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting [CD-ROM, p. 2572-2576]. Santa Monica, CA.

This work is not subject to U.S. copyright restrictions.