Assessing Performance in a Simulated Combat Information Center

John J. Lee, William L. Bewley, Barbara Jones, Hoky Min, and Taehoon Kang
UCLA/CRESST
Los Angeles, CA
johnjn@ucla.edu, {bewley, bjones}@cse.ucla.edu, hoky@ucla.edu, tkang@cse.ucla.edu

ABSTRACT

This paper describes research directed at determining the validity of measures of cognitive readiness, the mental preparation needed to be competent in the performance of complex tasks in a military environment. The Surface Warfare Officers School (SWOS) Multi-Mission Team Trainer (MMTT) is used to assess the performance of Tactical Action Officers (TAOs) operating in a simulated Combat Information Center (CIC). Scenarios require the TAO to defend against air, surface, and subsurface threats. A computer-based assessment system was developed for gathering data, analyzing, and reporting results. The system supports the assessor in rating the quality of learner responses to various scenario events on an "optimal," "adequate," and "other" rating scale, providing prompts for behaviors to record and questions to ask. The point and click interface minimizes interference with the assessor's observation of events and performance. The system automatically records and scores performance. Measures include (a) observed actions, e.g., reports appropriately communicated; (b) responses to mid- scenario probe questions, e.g., expectations regarding a track; (c) part-task anticipation requiring the learner to respond to short scenarios presenting a situation, e.g., identification of the greatest threat; and (d) critical events presenting cognitive "traps" designed to expose a cognitive error. Measures are mapped to cognitive constructs including situation awareness, decision making, communication, problem solving (formulating tactics plans), command and control (implementing and monitoring tactics plans) and acting effectively in a timely manner. The system aggregates results and generates graphs of performance by construct, identifies areas of strength, and provides recommendations for improvement. The paper describes the tool, the results of preliminary testing, the strengths and weaknesses of the approach to assessing cognitive readiness, and plans for future research.

ABOUT THE AUTHORS

John J. Lee is a Senior Researcher at the National Center for Research on Evaluation, Standards and Student Test (CRESST). Dr. Lee's current research is related to technology-based assessments in a variety of Navy/Marine Corps contexts. He is currently working on the development of a computer- based assessment tool for assessment of Tactical Action Officers (TAO) in a simulated CIC (Combat Information Center) onboard Navy ship called the Multi-Mission Team Trainer (MMTT). He is also working on a simulation- based re-certification assessment of marksmanship coaches' fault checking ability that delivers just-in-time, individualized instruction using Bayesian networks for diagnosis and remediation, and a game-based assessment project for the Navy related to assessment of complex skills (starting with damage control), also using Bayesian networks for real time and after action assessment of skills including situation awareness, decision making and communication.

William L. Bewley is an Assistant Director at CRESST. His experience combines research on human learning and performance, design and development of computer-based courseware, software product development, advanced technology applications for education and training, and program and product management. He manages all military projects at CRESST. His current work is concerned with cognitive readiness, assessment methodology, technology evaluation, and research on applications of advanced technology to human performance assessment, focusing on complex skills.

Barbara Jones is a research associate at CRESST and her work focused on best instructional and assessment practices for literacy and critical thinking, including developing literacy assessment frameworks, academic language performance measures, and corresponding instructional strategies to meet the learning needs of English Language Learners. Her most recent work with the Navy has included the assessment of TAOs (Tactical Action Officers) in the Multi-Mission Team Trainer (MMTT).

Hoky Min is a Graduate Student Researcher (GSR) at CRESST. His research areas include language acquisition, language assessment, and educational measurement. In particular, he has focused on tense usage in academic journals, differential item functioning of reading items for students with different academic backgrounds, and measurement invariance of English tests across English Language Learners (ELLs) with different language backgrounds. Hoky currently works on second language literacy development of ELLs.

Taehoon Kang is a Senior Researcher at CRESST. Dr. Kang's expertise is in the application of advanced psychometric methodologies to a number of measurement issues that have evolved through working with the educational and psychological tests. His research interests include the study of item response model selection, model-item fit assessment, differential item functioning, item parameter linking, and test equating. His current research focuses on development and applications of item-fit indices to find misfit items effectively under various testing conditions.

Assessing Performance in a Simulated Combat Information Center

John J. Lee, William L. Bewley, Barbara Jones, Hoky Min, and Taehoon Kang
UCLA/CRESST
Los Angeles, CA
johnjn@ucla.edu, {bewley, bjones}@cse.ucla.edu, hoky@ucla.edu, tkang@cse.ucla.edu

INTRODUCTION

Reduced manning and increased mission requirements require greater competencies of Naval warfighters than ever before. To support training in the required competencies, there is a need for high-quality approaches to performance measurement and assessment. Many of the competencies are part of the cognitive readiness model posited by Morrison and Fletcher (2002), who define cognitive readiness as:

> . . . . the mental preparation (including skills, knowledge, abilities, motivation, and personal dispositions) an individual needs to establish and sustain competent performance in the complex and unpredictable environment of modern military operations. (Morrison & Fletcher, 2002, p. 1-3).

One of the most important of the required cognitive readiness competencies is decision making in tactical environments. This is a focus of the Navy Surface Warfare Officers School (SWOS) Department Head (DH) course, the primary professional Surface Warfare Officer career course in the Navy's training continuum. A department head (DH) is an officer in charge of a ship department, e.g., Engineering, Operations Officer, Deck, Weapons, and Combat Systems. One of the most important roles played by all DHs is to stand watch in the Combat Information Center (CIC) as the Tactical Action Officer (TAO).

The TAO is responsible for tactical employment and defense of the ship. He or she manages use of the ship's weapons and sensors, directs the movements of the ship, and monitors the movements and actions of friendly and enemy ships, planes, missiles, and submarines in the region. The TAO must integrate this information to form a tactical picture of the situation, select appropriate responses, and issue orders. A TAO's responsibility is to determine the threat level of unknown tracks and to act appropriately in order to achieve the goals of the mission and adhere to the rules of engagement.

The SWOS DH course prepares officers to perform the duties of a TAO. One of the technologies used in TAO training is a simulated CIC called the Multi-Mission Team Trainer (MMTT). During instruction, students play the role of the TAO in several scenarios, with other students playing the roles of other CIC watchstanders. The other watchstanders sit at computer stations with microphones and headsets to monitor and provide information related to various warfare battle spaces (air, surface, undersea) to the TAO and outside the CIC to related personnel (e.g., helicopter or other air support). The MMTT is also used for assessment purposes.

SWOS has designed the MMTT to focus on the conceptual knowledge and skill of the TAO. Research has shown that simulations designed to assess conceptual knowledge and skill can be very effective when they require the cognitive demands of the underlying tasks (e.g., Psotka, Legree, Belanich, Bludau & Gray, n.d.; Virzi, Sokolov, & Karis, 1996). Beaubien and Baker, 2004, subdivide fidelity into three parts, environment fidelity, equipment fidelity, and psychological fidelity. Environment fidelity has to do with the match of the simulator with real situations based on motion cues, visual cues and other stimuli from the task environment. The MMTT's environment fidelity is high. Equipment fidelity refers to how well the simulator duplicates the appearance and feel of the real equipment. is Because it focuses on conceptual and knowledge and skill, not the specifics of operating equipment, the MMTT does not represent the displays and controls used on board different ships. Psychological fidelity is "the degree to which the trainee perceives the simulation to be a believable surrogate for the trained task. (Beaubin and Baker, 2004, p. i52)." This is difficult to determine for the MMTT, and will likely vary depending on the student, depending on ability to focus on the task rather than the training environment or equipment.

One of the issues related to the environment is the high cognitive load placed on students. A MMTT exercise can be high stress, with a large amount of information coming to the student at a rapid rate from multiple data sources. Not only must the student attend to the visual displays of tracks, he or she must also try to focus attention on communications coming through the headsets with different commands/messages being relayed in each ear. The student must be able to glean the necessary information, determine what is relevant, and then make decisions in a timely manner. For a track, the TAO needs to know where it is relative to the ship, its profile (direction, altitude, and speed), what the track has been doing, and recognize whether the track is a threat and if the threat could be attacked if necessary (Morrison, Marshall, Kelly and Moore, 1997).

Cues are of utmost importance when determining whether an unknown track is a friend or foe. For example, for air defense, Liebhaber, Kobus, and Feher (2002) identified the top six most important cues as origin, intelligence, IFF (identify friend or foe mode), airlane (published or known commercial air route), ES (electronic support used to detect threats via electromagnetic radiation), and maneuvers (e.g., the track is following the ship or maneuvering in specific ways). For surface warfare, the top three cues are: platform type, weapons envelope (the zone around the vessel's ship where they are within weapons range of the mission essential unit), and electronic emissions (Liebhaber & Feher, 2002). For undersea warfare, there are cues (acoustic and non-acoustic) related to submarine locations (or possible locations, known as "datums"), torpedo danger zones, and intelligence gathered from other sensors or communications. Non-acoustic signatures include a submarine's magnetic and electrical signatures and the submarine's wake (Naval Doctrine Command, 1998).

There is also the issue of emotional load (Menaker, Coleman, Collins, & Murawski, 2006), which is related to the psychological fidelity of the simulation. If students are able to suspend disbelief and feel the high-level emotions of the battle space, the high-stakes nature of the final exams (counting toward their final grades) and time stress, the MMTT may provide a place for students to practice emotional control.

The goal of this research is to test methods for assessing TAO performance in the MMTT. Currently, TAO performance assessment is based on whether or not certain actions occur, e.g., did the TAO order queries and warnings to be sent to suspected hostile tracks, did the TAO send an airplane to visually identify a suspected hostile track, or did the TAO shoot at a threatening hostile track? This is the "what" of performance in the MMTT. In addition to these measures, SWOS is concerned with measuring the TAO's cognitive readiness or the "why" of performance That is, in addition to asking whether or not certain actions were performed, SWOS wants to know *why* they were performed—the thinking behind the actions. To do this, information is needed on the thinking supporting the judgments leading to actions.

This paper describes our approach and progress toward developing an assessment of TAO performance that includes both the what and the why, and the results of initial steps toward validation of the assessment and standards setting for grading purposes required by SWOS. We begin with a description of our approach, including the method used to assess the thinking behind actions and definition of the constructs to be measured. This is followed by a description of the method used to conduct initial studies of reliability and validity of the assessment. We close with a description of preliminary results and a summary and discussion.

APPPROACH

As suggested by Tenney & Pew (2006) , Jones & Endsley (2004) and many others, there are several ways to assess the "why," or the thinking behind actions:

1. Mid-scenario probe questions. One can pose questions during the scenario without pausing the activity. The advantage is a low memory requirement because the report is immediate, not delayed. The disadvantages are that the questions may interfere with performance due to interruption, and they may direct attention to things that might be otherwise overlooked.
2. Part-task anticipation. Questions are posed following short scenarios or scenario fragments presenting a situation. The advantages are that the assessment is brief and focused and the memory requirement is low. The disadvantage is that short trials are not representative of real situations and workloads.
3. Critical events. An event presenting a problem or anomalous situation is inserted in the scenario, and the assessor observes the response. This is the approach taken by Smith-Jentsch, Johnston, and Payne (1998) in an earlier study of TAO tactical decision making, and more recently by Radtke, Johnston, Biddle, and Carolan (2007) in a study of pilot decision making in air combat. The advantages are a low memory requirement and no mid-scenario questions to intrude on the student's task performance. The disadvantages are that the critical events elicit behaviors, the

"what" of performance, but do not provide information on the "why," the thinking behind the actions.

4. After-action review. This is the usual approach to assessment in team situations. The assessor and team review and discuss performance at the completion of the simulation scenario. The advantage is that it's not intrusive. The disadvantage is a very high memory requirement for students and assessors, although performance data collected during the scenario can be used to support the AAR, e.g., as done in the critical event study by Radtke, et al. (2007).

The approach taken in this research is a blend of all four, designed to eliminate or minimize the disadvantages while retaining the advantages of each approach separately. We looked at the total scenario as a string of connected part-tasks, with probe questions at natural pauses between parts, and we included critical events in each part-task designed to elicit not only actions, but actions that expose "cognitive errors," mistakes attributable to errors in the TAO's critical thinking. Data on cognitive errors can be used to group errors in cognitive bias categories described by Groopman (2007) for medical decision making and by several studies of errors in intelligence analysis, e.g., CIA Directorate of Intelligence (1997), George (2004), and Heuer (1999). In addition to providing information on critical thinking, the approach reduces the interruption due to questions and the memory requirement by asking questions only during the pauses and by timing pauses so that there is little delay between the response and the question. It also avoids the risk of the question cueing different behavior because the questions are asked after the event and response have occurred. And because the scenario fragments are strung together to form a long scenario, it avoids the problem of testing with unrepresentative situations and workloads.

Measures and Constructs

We defined the constructs to measure based on the cognitive demands required to successfully complete the tasks in the simulation. We based our construct list on the constructs identified by Morrison and Fletcher (2002),prior work with tactical decision making (Radtke, Johnston, Biddle, & Carolan, 2007; Morrison, Marshall, Kelly, & Moore, 1997), and interviews with SWOS staff. TAOs need to be able to use their situation awareness to establish a rapid picture of the environment, which includes determining if a track poses a threat, making a decision to collect more information, or responding to an identified threat as dictated by standard operating procedures, all while adhering to the rules of engagement (ROE). The constructs we defined included: 1) Situation awareness, 2) Problem solving/ Tactics planning 3) Decision making, 4) Tactics plan implementation and monitoring, 5) Acting effectively in a timely manner, and 6) Communication. The first three constructs were grouped under a more general construct called Thinking, and the last three constructs were grouped under a more general construct called Action.

Situation Awareness

Endsley (1988; 1997; 2000) defines situation awareness (SA) as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future." Her definition delineates three levels of situation awareness: Level 1, Perception of elements in the environment; Level 2, Comprehension of the current situation; and Level 3, Projection of future status. They are listed in increasing degree of cognitive demand.

The three levels relate to the following questions:

- Level 1: What is going on? What elements in the environment should you attend to? What elements are relevant (critical cues) for the given situation?
- Level 2: Do you know why the relevant cues are important? Which are not and why? What patterns do you see?
- Level 3: What are you expecting to happen?

In the MMTT simulation, Level 1 SA refers to choosing which variables to consider, focusing on the correct cues, and being sure to consider a wide range of variables. Level 2 SA is understanding why those variables are important and how they relate to each other and to an interpretation. Level 3 SA is being able to predict what might occur in the future in order to help you determine what course of action to take.

Problem Solving: Formulating tactics plans

The problem solving phase involves being able to transform the mission goal and subgoals into plans (Morrison & Fletcher, 2002). The TAO must use situation awareness to determine potential courses of action (COA) that can be undertaken. The goal in the CIC is tied to the primary mission of protecting the mission-essential unit, the ship most essential to successful completion of the mission. The tactics plan will involve utilizing resources (data gathering through sensors and real assets like helicopters and airplanes), and considering the reaction of the unknown track to friendly force action (e.g., responding or not to queries and warnings) in order to develop defensive and/or offensive action plans based on the rules of engagement. Sometimes, when the situation is recognized as matching a situation experienced before, the TAO will not have to develop a plan (Klein, 1999). Rather, he or she can select a plan used successfully in the prior situation, or if the situation is similar but not identical to a situation experienced before, the TAO can edit the plan used successfully in the prior situation and test it by running a mental simulation.

Decision Making

Decision making is the process of choosing from among alternative plans. As noted above, if the situation is recognized, the associated successful plan will be selected, or a plan associated with a similar situation will be edited and tested in a mental simulation (Klein, 1999). If the situation is not recognized, however, plans will be developed and alternative plans will have to be evaluated and one of them chosen. This may involve using some mental simulations or "what if" thinking.

Implementation and Monitoring of Tactics Plans (Command and Control)

Effective implementation involves making sure that the chosen COA is carried out as planned, and to maintain or revise the plan based on new information from the results of the implementation. If the situation takes a turn for the worse, then a contigency, modified or alternative plan must be put into place. In the MMTT, the orders given by the TAO and queries to the appropriate watchstanders along with the monitoring of the situation are relevant indicators of this construct.

Communication

This construct deals with the proper communication of the TAO with others, including other watchstanders and superiors. It includes having the correct structure (following protocol and having the right format), treatment (message content, accuracy, delivery style/ presentation), and timeliness. Researchers (Achille, Schulze, Gladwell, and Schmidt-Nielsen, 1995; Urban, Bowers, Monday and Morgan, 1995) describe four aspects of effective communication, particularly in military environments:

1. Accuracy: unambiguous and proper use of terms.
2. Terseness (brief): especially important in heavy use communication nets
3. Selective (relevance): only pertinent information.
4. Identification (to guide attention): communicator identifies himself/herself by name and/or by role

The importance of effective communication cannot be overemphasized, given issues related to message interference due to overload of the communications network, messages "stepping on" each other, and the difficulty of handling multiple and different communications coming into different sides of the TAO's and others' headsets.

Acting Effectively in a Timely Manner

Another important skill for a TAO is the ability to act effectively in a timely manner. Just performing the correct action is not enough; it has to be done within a certain amount of time, known as the window of opportunity (Cothier and Levis, 1986). In the MMTT, this means being more proactive than reactive, for in the latter case, the may be too late. Since this is a high stakes situation, being able to act effectively can make a large difference in the outcome of the mission.

METHOD

Subjects

165 Department Head students participated in the study from three DH classes. The classes had 58, 47, and 51 students. Most were ranked lieutenants (91.7%), and most had 5-10 years of service in the Navy (69.5%). A majority of the students were also most recently deployed to the Middle East (53.8%), with the next highest percentage deployed to the Western Pacific (14.1%). All but one (with six missing cases) had their deployment end within the last seven years, and most (86.6%) said their deployment supported ongoing operations. About 91% of the students had completed some Combat Systems Officer (CSO) training, and 81% either this year (29%) or last (52%). 134 students were CICWO (Combat Information Center Watch Officer) qualified, with 72% qualified in 2002 or 2003 (about six or seven years ago). A small percentage (18.6%) of students were actually already TAO qualified, most (72.4%) in 2005 and 2006.

Tasks

Each DH student performed as a Tactical Action Officer (TAO) on one of two final scenarios. A division officer played the roles of all the other watchstanders. Both scenarios take place in littoral environments, a compressed battle space, although the second scenario is more compressed than the first.

The scenarios lasted between about 30-45 minutes and involved events including air, surface and subsurface contacts being addressed in varied order, including some simultaneous events.

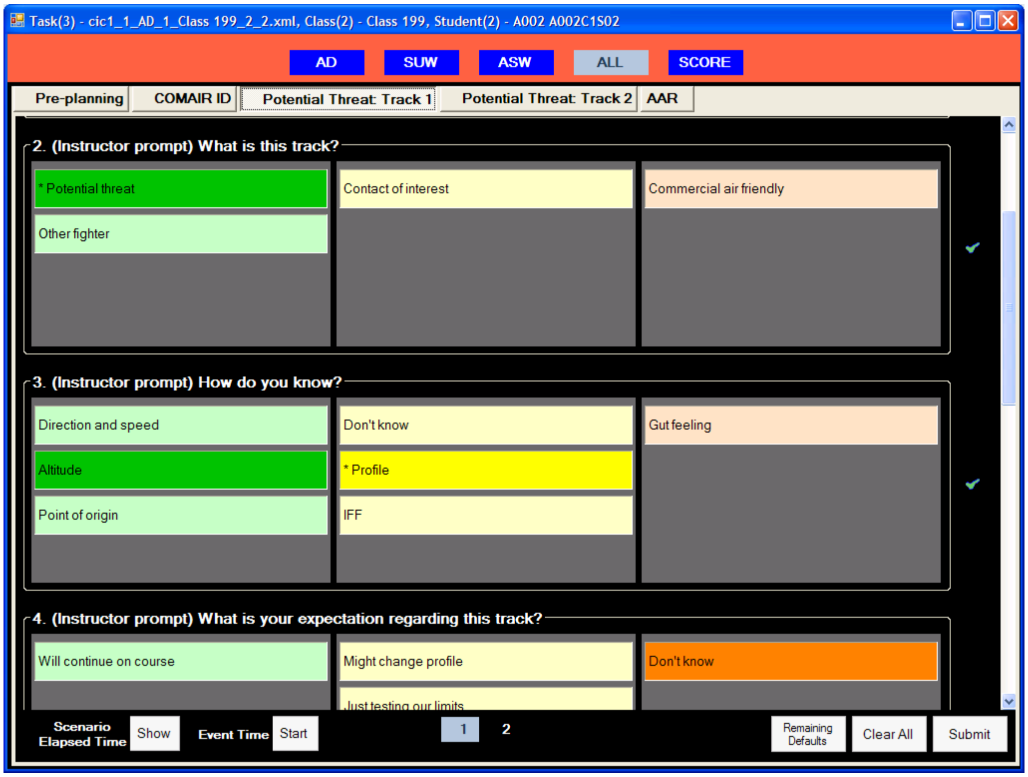
The Assessment Tool

The assessment tool is a PC-based program providing a series of screens mapped to the events of the MMTT scenario. It is used to assess the performance on the two MMTT finals as a criterion referenced summative test. Figure 1 shows a screen for an air defense scenario. Each screen contains items (statements or questions) linked to several descriptions of student actions or responses. In the figure, the items are:

- Queries and warnings
- 1.What is this track?
- 2. How do you know?
- 3.What is your expectations regarding this track?

The descriptions of student actions or responses associated with each item are shown as buttons grouped into three categories:

- Optimal, the green buttons at the left
- Adequate, the yellow buttons in the middle
- Other, the red buttons at the right



Task(3) - cic1_1_AD_1_Class 199_2_2.xml, Class(2) - Class 199, Student(2) - A002 A002C1S02

2. (Instructor prompt) What is this track?

* Potential threat	Contact of interest	Commercial air friendly
Other fighter		

3. (Instructor prompt) How do you know?

Direction and speed	Don't know	Gut feeling
Altitude	* Profile	
Point of origin	IFF	

4. (Instructor prompt) What is your expectation regarding this track?

Will continue on course	Might change profile	Don't know
	Just testing our limits	

Figure 1. An Assessment Tool screen.

Descriptions and groupings into categories were initially defined by one rater and then modified with assistance from six expert raters. The rater records the student's responses and actions for each item by selecting the appropriate buttons. When the rater clicks the "Submit" button at the lower right corner of the screen, the items are scored and the next screen is displayed. Buttons at the top of the screen (AD, SUW, ASW, All, and Score) are used to filter items, e.g., only air defense (AD), surface warfare (SUW), or anti-submarine warfare (ASW) item, show all items, or view the score report of student performance. The tabs below these buttons show scenario events in chronological order, left to right. The rater moves from event to event by selecting the tabs. Items that require the evaluator to prompt the student by asking a question verbally have a "instructor prompt" written in parentheses.

Procedure

DH students took one of two alternative test scenarios for the TAO MMTT final examination. Each student sat at a PC emulating a TAO workstation. All voice messages were delivered through headphones, with different channels for each headphone, just as experienced by the TAO in a real CIC. One person played the role of all the watchstanders in the CIC, and triggered MMTT events defined by the scenario and based on the students actions, e.g., issuing queries and warnings as directed by the student. Students were given a pre-brief of the test situation the day before. During the examination, a rater sat next to the student and recorded performance with the Assessment Tool using a laptop PC. The rater recorded all student actions/responses for each item Each scenario includes a pre-planning tab, which asks the student what they perceive as being the important components of the mission (e.g., protect the mission essential unit), and what they are expecting to happen. The scenario is then started. Students use their computer displays to look at sensor data (including radars) and use voice communications both internally and externally using a communications interface and headset. As the end of the scenario, students are given a set of after action review questions and the rater uses the Assessment Tool score report to provide the student feedback on performance (both good and bad) and the pass/fail decision.

In addition to the total score, there is a tab in the assessment tool that keeps track of single point failures, errors that a student makes in the scenario that are costly in terms of lives, assets, or political relationships. For example, breaking rules of engagement, running a ship aground, or friendly fire would fit into this category. If a student received one or more of "other" category in the single point failures tab, then the student failed the exam, regardless of their score on the rest of the items.

Scoring

A point value of 2 was given for each "Optimal" , 1 for "Adequate" responses, and -1 was given for the "Other" responses. Scores for all the responses for an item were summed to obtain the item score, and the item scores were used as the basis for further analyses. Construct scores were computed by summing the scores for the items mapped to each construct, and because possible scores were different for each construct, the scores were standardized by converting them to percent of the highest possible score as defined by expert raters.

Scenario Complexity

In order to determine relative complexity between the two MMTT finals, we asked eight DH instructors to rate the relative complexity of the 2 scenarios to rank order the relative level of difficulty. The order of scenarios showed that with high reliability (phi coefficient = .99 for eight raters and phi coefficient= 0.95 for two raters for all 23 scenarios) that the two finals were among the most difficult used in the MMTT, and that the second scenario was slightly more difficult than the first. The complexity scores (following the method used by Crane, Robbins, Bennett & Bell, 2001), which are transformed z-scores of the normally distributed proportions of judgments (each mean rank divided by the total number of scenarios minus one), are 3.73 for final scenario 1 and 3.78 for final scenario 2. The range was from 1.00 for the easiest surface warfare (SUW) scenario teaching voice communications to 4.07 for the compressed battlespace scenario.

Mapping Observed Behaviors to Constructs

We mapped the observed behaviors as recorded in the assessment tool to the constructs. For example, for situation awareness, we mapped the items that dealt with establishing the picture (what are the contacts of interest, what is the track?) to Level 1 SA. Items using questions like "how do you know?" to Level 2 SA and question like "what do you expect to happen?" to Level 3 SA.

We then conducted confirmatory factor analysis to see how many factors were found for each final. Factor analysis is a method for identifying the "factors, or underlying dimensions, that underlie the relations among a set of observed variables. (Pedhazur & Schmelkin, 1991, p. 66)".

Cut score determination and setting performance levels

SWOS requested that we determine the cut scores and performance levels for each MMTT final. To do this, we needed to determine the minimum and maximum expected scores for each final. Five subject matter experts were asked to enter results into the assessment tool based on a maximum score (what the best student might do) and based on a minimum competency level of performance (what they would consider just barely passing) for each of the MMTT finals.

Generalizability studies (G-studies; Shavelson & Webb, 1991) indicating how well measures taken in one context or environment generalize to another were conducted to determine if the raters were reliably scoring the minimum and maximum expected performances. A G-study and subsequent D (decision) study were conducted to determine how many raters were needed to obtain reliable results (VanLeeuwen, 1997). The G-study results showed phi coefficients of .78 (minimum score; one expert's score was dropped) and .97 (maximum score) for Final 1 and .83 (minimum) and .94 (maximum) for Final 2 to determine the number of raters needed for consistent expected minimum and maximum scores. The D-study results showed that five raters were sufficient to reliably measure the expected minimum and maximum scores used for the cut score determinations.

A modified Angoff Method (Zieky& Perie, 2004; Maurer, Alexander, Callahan, Bailey, & Dambrot, 1991) was used to determine each cut score. Each of five evaluators (panelists) provided two values for each item:

- The expected item-score for an imaginary student with minimal competency (*EIS*)
- The possible maximum item score (*MIS*)

When there are *I* items (*i=1,…,I*) and *P* panelists (*p=1,…P*), the cut-score was determined with the equation,

$$Cut - Score = \frac{ave(EIS)}{ave(MIS)} = \left(\frac{\sum_{p=1}^{P} \sum_{i=1}^{I} EIS_{pi}}{P \times I} \right) \div \left(\frac{\sum_{p=1}^{P} \sum_{i=1}^{I} MIS_{pi}}{P \times I} \right) \quad (1)$$

The possible cut-score ranges between 0 and 1. When the test is administered, the test score (*TS*) for each examinee (*j=1,…J*; his/her item *i*'s raw score is IRS_{ji}) is calculated as:

$$TS_j = \sum_{i=1}^{I} \frac{IRS_{ji}}{ave(MIS_i)} \times \frac{1}{I},$$

where $ave(MIS_i)$ for item *i* is $\dfrac{\sum_{p=1}^{P} MIS_{pi}}{P}$.

A table was then generated to show the cut score and six performance bands (one for failing and five more for passing through advanced) above it using equal intervals. This process was repeated for each final.

RESULTS

Data for three DH classes collected by SWOS using the Assessment Tool were sent to us for analysis. There were a total of 86 students who received the first test scenario Final 1and 79 received the second (Final 2). Students who did not pass on the first examination, were retested a day later with the alternative scenario. There are two scores for these students, one for Final 1 and one for Final 2. Across the two finals, there were only three students who failed the first time, one on Final 1 and two on Final 2. Failed scores were removed from the analysis, and no one failed on their second attempt.

Descriptive statistics are shown in Table 1. The scores for each final were normally distributed. The average for Final 1 was 139.68 out of a possible 178 (*SD*=17.77) and for Final 2, 93.80 out of a possible 115 (*SD*=12.81). The percent of the mean to the maximum was close to 80 percent for each final (78.5% for Final 1 and 81.6% for Final 2), and they are not significantly different.

Table 1. Descriptive Statistics for Two MMTT Finals

	N	Range	Min	Max	Mean	SD
Final 1	85	98	80	178	139.68	17.77
Final 2	80	58	62	115	93.80	12.81

Preliminary Validation Analysis

The validation approach we took was to conduct confirmatory factor analyses on the final scenarios to look for evidence of construct validity. We also looked for evidence of convergent validity by looking at the relationship of performance on the MMTT finals with performance on the TAO written final exam.

Constructs

As noted in the Methods section, six first-order constructs were identified on the basis of needs analysis on the abilities or knowledge required for TAO performance. The six constructs are: situation awareness level 2, situation awareness level 3, decision making, problem solving, tactics implementation and monitoring, timeliness and communication.

Technical Quality

The technical quality of the measures was partially determined by examining their construct validity and reliability.

*Construct validity***:** Construct validity refers to the degree to which assessment results can be interpreted as a meaningful measure of the construct we intend to measure. Thus, the supporting evidence for the construct validity for the MMTT finals would indicate that the test scores are meaningful indicators of the TAO constructs we intend to measure. In order to provide evidence for the construct validity of the MMTT finals, we ran a series of confirmatory factor analyses (CFA) for each of the constructs separately, using EQS 6.1 (Bentler, 2007). The robust maximum likelihood (ML) estimation method, Satorra-Bentler statistic was used for testing model fit and parameters because the MMTT measures are not distributed multivariate-normal (Satorra & Bentler, 1994). The number of items for constructs is found in Table 2. As can be seen, it ranges from three to nine items per construct.

Table 2. Goodness of Fit Summary

Model	X^2	*p*	*df*	CFI	RMSEA	RMSEA .90 CI	α
Final 1							
Situation Awareness Level 2 (9 items)	30.95	0.19	25	0.89	0.05	0.00-0.11	0.725
Situation Awareness Level 3 (7 items)	19.54	0.11	13	0.95	0.08	0.00-0.14	0.793
Decision Making (7 items)	10.48	0.57	12	1.00	0.00	0.00-0.10	0.704
Tactics Plan Implementation (6 items)	9.47	0.22	7	0.90	0.07	0.00-0.15	0.823
Communication (6 items)	8.82	0.27	7	0.99	0.06	0.00-0.15	0.746
Timeleness (5 items)	7.26	0.20	5	0.96	0.07	0.00-0.18	0.642
Final 2							
Situation Awareness Level 2 (5 items)	6.36	0.27	5	0.92	0.06	0.00-0.17	0.625
Situation Awareness Level 3 (3 items)		Not Applicable (Saturated Model)					0.336
Decision Making (3 items)		Not Applicable (Saturated Model)					0.640
Problem Solving (5 items)	3.26	0.66	5.00	1.00	0.00	0.00-0.12	0.519
Tactics Plan Implementation (6 items)	13.71	0.09	8.00	0.95	0.10	0.00-0.18	0.832
Communication (6 items)	33.36	0.12	25	0.92	0.07	0.00-0.12	0.713

* Maximum Likelihood estimation method with Satorra Bentler scaled Chi-square (corrected for normality)

Table 2 also displays the fit of the CFA models for each construct. The model fits indicate the extent to which Model fit represents the degree to which the researcher's hypotheses or theories correctly reflect the data and it is evaluated by various types of fit indexes such as the model chi-square (X^2), root mean square of error approximation (RMSEA;Steiger, 1990), comparative fit index (CFI; Bentler, 1990). In general, the CFA models exhibited good fit in that X^2 is not significant, CFI is larger than 0.90, RMSEA is smaller than 0.10. Model fit was not obtained for Situation Awareness 3 and Decision Making for Final 2 because the model is saturated (the model-implied covariances are identical with the observed covariances). Also, all of the factor loadings in the CFA models were statistically significant at the level of 0.05. The good model fit and significant factor loadings provided the evidence that the MMTT measures are meaningful indicators of the TAO constructs intended.

Construct Scoring

When there are multi-level relationships among constructs, authors have cautioned the use of subdomain level (construct) level scores to compare individuals with each other (de la Torre & Song, in press). We computed average scores for individual students by construct so that any given student could see how they performed relative to the constructs.

Reliability of Construct Scores

Reliability refers to the consistency of assessment results. Low reliability indicates that the scores are influenced by random errors such as the level of students' motivation or the inconsistency in rater scoring across students. Reliability also constitutes a necessary condition for construct validity. Cronbach alpha (Pedhazur & Schmelkin, 1991) was obtained as an estimate of the reliability for each of the MMTT finals. The last column in Table 2 shows the alpha reliabilities by construct. Most of them are relatively high, with the exception of Situation Awareness Level 3 and Problem Solving on Final 2. More data is needed to gain better estimates of these reliability coefficients.

Convergent Validity: Correlations with other background variables and measures

If the MMTT assessment measures are valid, they should correlate with other background variables and measures thought to be related to the constructs. This is referred to as convergent validity. We looked at the correlation between background variables and their MMTT Final scores. For MMTT Final 1, there were no significant correlations. However for MMTT Final 2, there were significant correlations with two qualifications: CIC Watch Officer (CICWO) Qualified, *r(72)*=.233, *p*=.047, and TAO qualified, *r(72)*=.346, *p*=.003. In other words, the longer it has been since you got CICWO or TAO qualified, the better you performed on the MMTT Final 2.

We also looked at the correlation between the scores on the MMTT Finals and the TAO Class Final Exam scores.

> There are several versions of the TAO Class Final Exam. Each exam has a scene setter that is given to the student a few days ahead of the exam. The exams are scenario driven and cover all of the warfare areas (AD/IO-Air Defense/Information Operations, SUW-Surface Warfare, USW-Undersea Warfare and EXW-expeditionary warfare). They are designed to test the student's understanding of the specific tactics, threat prioritization and tactical decision process. Responses are open-ended paragraphs. (G.Chapman, personal communication, 2009).

As shown in Table 3, the correlation between the MMTT Final 1 score and the TAO Final Exam overall score was significant (*r*=.261, *p*=.021). The correlation between the the MMTT Final 2 score and the TAO Final Exam overall score was not significant (*r*= 0.12, *p*=.287).

Table 3. Correlations between MMTT Scores and TAO Final Exam Scores

MMTT Final	TAO Final exam
Final 1 (n= 78)	.261*
	p=.021
Final 2 (n=78)	.122
	p=.287

SUMMARY AND DISCUSSION

In this paper we have described the development of measures of cognitive readiness in the performance of Tactical Action Officer skills in the Multi-Mission Team Trainer (MMTT), a simulated Combat Information Center used in the Surface Warfare Officers School Department Head course.

The process required having experts give us maximum and minimum passing competencies, then determining a cut score and five more performance levels above the cut score for each MMTT final. Generalizability analyses were conducted to determine the reliability of those ratings. The assessments show evidence of both reliability and construct validity. Confirmatory factor analysis revealed six constructs for each final, though slightly different. The MMTT Final 1 had mapped to it Level 2 situation awareness, Level 3 situation awareness, timeliness, decision making, tactics plan implementation and communication. The MMTT Final 2 had mapped to the same constructs with the exception of timeliness, and instead had mapped to the construct of problem solving. This may be due to the Final 2 scenario requiring more problem solving skills because the TAO may not have a pre-planned response to the simultaneous events occurring below the surface of the water, on the surface and in the air. Scenario complexity was shown to vary across the 23 scenarios used in the MMTT training and assessments. Further validity evidence is needed to validate the measures. This includes data from more students.

There is interest in using a similar assessment in the fleet context. An authoring tool has been developed for adding future scenarios. Future work includes incorporating Bayesian networks to provide diagnosis and remediation capabilities for using the tool for instructional purposes as well.

ACKNOWLEDGEMENTS

This research was sponsored by Office of Naval Research Grant No. N000140810126. We thank RADM (Sel) Jeffery S. Jones, CDR Winton Smith, CAPT Carol Hottenrott, Mr. George Ponsolle, CDR Karl Tenney, LCDR Matthew Jerbi, CDR Greg Chapman, LCDR Brad Hawksworth, and the SWOS Department Head students who participated in this study. We also wish to thank David Monroe, Hongwen Cai, who assisted in the work, and our bird dog, Dr. Susan Coleman, Intelligent Decision Systems, Inc., for her oversight and valuable suggestions.

REFERENCES

Achille, L.B., Schulze, K.G., & Schmidt-Nielsen, A. (1995). An analysis of communication and the use of military terms in Navy team training. *Military Psychology: Special issue: Team processes, training, and performance, 7*, 95-107.

Beaubin, J.M., & Baker, D.P. (2004). The use of simulation for training teamwork skills in health care: how low can you go? *Quality Safe Health Care, 13(i51-i56).*

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychol Bull, 107(2), 238-246.

Bentler, P. M. (2007). EQS 6.1 for Windows[Computer Software]. Encino: CA Multivariate Software.

CIA Directorate of Intelligence. (1997). A compendium of analytic tradecraft notes: Volume I, notes 1-10. Retrieved November 29, 2006 from http://www.au.af.mil/au/awc/awcgate/cia/tradecraft_notes/contents.htm#contents.

Cothier, P.H., & Levis, A.H. (1986). Timeliness and measures of effectiveness in command and control. *IEEE Transactions on Systems, Man, and Cybernetics,16*(6), 844-853.

Crane, P., Robbins, R., Bennett, W.B., & Bell, H.H. (2001). Mission complexity scoring for distributed mission training. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

de la Torre, J., & Song, H. (in press). Simultaneous estimation of overall and domain abilities: A higher-order IRT model Approach. *Applied Psychological Measurement.*

Endsley, M.R. (1988). Design and evaluation for situation awareness enhancement. *In Proceedings of the Human Factors Society 32nd Annual Meeting, 97-101*. Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1997). The role of situation awareness in naturalistic decision making. In Zsambok, C. E. & Klein, G. (Eds.), *Naturalistic Decision Making* (pp. 269-283). New Jersey: Lawrence Erlbaum Associates.

Endsley, M.R. (2000). Theoretical underpinnings of situation awareness: A critical review. In MR. Endsley and D.J. Garland (Eds.), Situation awareness analysis and measurement (pp. 3-32). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

George, R. Z. (2004.) Fixing the problem of analytical mind-sets: Alternative analysis. International Journal of Intelligence and Counterintelligence, 17, 385-404.

Groopman, J. (2007.) How doctors think. Boston: Houghton Mifflin Co.

Heuer, R. J. (1999.) Psychology of intelligence analysis. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

Jones, D.G., & Endsley, M.R. (2004). Use of real-time probes for measuring situation awareness. *The International Journal of Aviation Psychology, 14*(4), 343-367.

Klein, G. (1999). Sources of Power: How People Make Decisions. The MIT Press.

Liebhaber, M.J., Kobus, D.A., & Feher, B.A. (2002). *Studies of U.S. Navy air defense threat assessment: Cues, information order, and impact of conflicting data* (Tech. Rep. SSC-1888), San Diego, CA: Space and Naval Warfare Systems Center.

Liehbhaber, M.J., & Feher, B.A. (2002). *Surface warfare threat assessment: Requirements definition* (Tech. Rep. SSC-1887). San Diego, CA: Space and Naval Warfare Systems Center.

Maurer, T.J., Alexander, R.A., Callahan, C.M., Bailey, J.J., & Dambrot, F.H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel psychology*, *44*, 235-261.

Menaker, E., Coleman, S., Collins, J. & Murawski, M. (2006). *Harnessing, experiential learning theory to achieve warfighting excellence.* Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

Morrison, J.E., & Fletcher, J.D. (2002). *Cognitive readiness.* IDA Paper P-3735. Institute for Defense Analyses.

Morrison, J.G., Marshall, S.P., Kelly, R.T., & Moore, R.A. (1997). *Eye tracking in tactical decision making environments: Implications for decision support evaluation*. Proceedings of the Third International Command and Control Research and Technology Symposium, National Defense University.

Naval Doctrine Command (1998). *Littoral anti-submarine warfare concept.* Retrieved January 12, 2009, from http://www.fas.org/man/dod-101/sys/ship/docs/ aswcncpt.htm.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach (student edition)* . Lawrence Erlbaum.

Psotka, J.P. , Legree, P., Belanich, J., Bludau, T.M., & Gray, D. (n.d.). *New simulator-based training approaches for security operations: Low-fidelity simluations for assessment*. Unpublished manuscript.

Radtke, P., Johnston, J. H., Biddle, E., & Carolan, T.F. (2007, December). Integrating and presenting performance information in simulation-based air warfare scenarios. Proceedings of the interservice/industry training, simulation and education conference, Orlando, FL.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. v. Eye & C. C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp. 399-419). Thousand Oaks, CA: Sage.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smith-Jentsch, K. A., Johnston, J. H., & Payne, S. C. (1998.) Measuring team-related expertise in complex environments. (pp. 61-87.) In J. A. Cannon-Bowers & E. Salas (Eds.), Making decisions under stress. Washington, DC: American Psychological Association.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. Multivariate Behavioral Research, 25(2), 173-180.

Swanson, D.B., Norman, G.R., & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. *Education Researcher*, *24, 5-11,35.*

Tenney, Y. J., & Pew, R. W. (2006).Situation awareness catches on: What? So what? Now what? (pp. 1-34). In R. C. Williges (Ed.), Reviews of human factors and ergonomics, Volume 2. Santa Monica, CA: Human Factors and Ergonomics Society.

Urban, J.M., Bower, C.A., Monday, S.D., & Morgan, Jr., B.B. (1995). Workload, team structure and communication in team performance. *Military Psychology*, *7*(2), 123-139.

VanLeeuwen, D.M. (1997). Assessing reliability of measurements with generalizability theory: An application to inter-rater reliability. *Journal of Agricultural Education,38*(3), 36-42.

Virzi, R.A., Sokolov, J.L., & Karis, D. (1996). *Usability problem identification using both low- and high-fidelity prototypes.* In Proceedings of Human Factors in Computing Systems: CHI ’96 (pp.236-243). New York: ACM.

Zieky, M., & Perie, M. (2004). *A primer for setting cut scores on tests of educational achievement.* Retrieved March 3, 2009 from http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf. Princeton: Educational Testing Service.