REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-06-2006	Final	01-05-2005 – 30-06-2006

4. TITLE AND SUBTITLE
Development of a SWOS Assessment System Concept

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-05-1-0659

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Eva L. Baker, William L. Bewley, Gregory K. W. K. Chung, and Girlie C. Delacruz

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

UCLA CSE/CRESST
300 Charles E. Young Drive North
GSE&IS Bldg. 3rd Flr./Mailbox 951522
Los Angeles, CA 90095-1522

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 North Randolph St.
Arlington VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The UCLA Center for Research on Evaluation, Standards, and Student Testing (UCLA/CRESST) and CRESST's subcontractor, the University of Southern California Center for Cognitive Technology (USC/CCT), developed and demonstrated a Navy Surface Warfare Officers School (SWOS) assessment system concept by (a) defining functions and a concept of operation for the assessment system, and (b) developing an initial prototype of a component of the system to illustrate the concept and its potential. The prototype was based on the air defense planning task of the SWOS Department Head course. Assessments included a simulation that attempted to mimic the actual planning task, background surveys including measures of prior air defense experience, measures of conceptual knowledge including a survey of threat assessment knowledge and a knowledge mapping assessment of threat detection-to-engage process knowledge, and self-ratings of air defense knowledge.

15. SUBJECT TERMS
Assessment system, air defense planning training

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified			

Final Report

Deliverable - June 2006

Development of a SWOS Assessment System Concept

Eva L. Baker
CRESST/University of California, Los Angeles

Office of Naval Research
Award # N00014-05-1-0659

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
301 GSE&IS, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under Office of Naval Research Award Number #N00014-05-1-0659, as administered by the Office of Naval Research. The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the Office of Naval Research.

TABLE OF CONTENTS

FINAL REPORT

William L. Bewley, Gregory K. W. K. Chung, and Girlie C. Delacruz

CRESST/University of California, Los Angeles

Allen Munro

Center for Cognitive Technology/University of Southern California

Abstract

The UCLA Center for Research on Evaluation, Standards, and Student Testing (UCLA/CRESST) and CRESST's subcontractor, the University of Southern California Center for Cognitive Technology (USC/CCT), developed and demonstrated a Navy Surface Warfare Officers School (SWOS) assessment system concept by (a) defining functions and a concept of operation for the assessment system, and (b) developing an initial prototype of a component of the system to illustrate the concept and its potential. The prototype was based on the air defense planning task of the SWOS Department Head course. Assessments included a simulation that attempted to mimic the actual planning task, background surveys including measures of prior air defense experience, measures of conceptual knowledge including a survey of threat assessment knowledge and a knowledge mapping assessment of threat detection-to-engage process knowledge, and self-ratings of air defense knowledge. Pilot test results were mixed, with weak but inconsistent evidence of an expected positive effect of air defense experience, sensitivity of measures to instruction, and intercorrelations among presumably related measures. Results provided information leading to needed revisions of assessment content and assessment tool user interfaces, however, and preliminary results of a subsequent test suggest that the revisions have led to improved results. Next steps are described, including research on differences among simulation scenarios and the use of assessment results for performance predictions supporting placement and development decisions.

Introduction

In order to improve warrior proficiency, the Navy Surface Warfare Officers School (SWOS) plans to transform Department Head preparation to make it more responsive to the training needs of the individual by recasting training to be a hybrid self-paced and group-paced curriculum enhanced with ample practice in tactical modeling and simulation for proficiency development. Modifying some instructional modules to enable self-pacing will allow the Navy to take advantage of students' experience base to capture time that would have been spent in unnecessary training. This captured time will be repurposed to provide students with additional practice in tactical skill development through modeling and simulation.

To execute this plan, SWOS requires a reliable, valid, and integrated system of assessments that includes initial assessments for placement purposes, formative assessments for development purposes, and exit assessments for certification of proficiency purposes. The UCLA Center for Research on Evaluation, Standards, and Student Testing (UCLA/CRESST) has the experience and technical expertise in designing assessment systems to support SWOS in developing the required assessment system, and the Department Head course provides an important application area and testbed for developing assessments supporting the Office of Naval Research (ONR) Future Naval Capability (FNC) initiative.

To support SWOS while pursuing FNC objectives, ONR funded UCLA/CRESST to develop and demonstrate a SWOS assessment system concept by (a) defining functions and a concept of operation for the assessment system, and (b) developing an initial prototype of a component of the system to illustrate the concept and its potential.

The Navy, like many organizations requiring training in tasks requiring high-level cognitive skills, relies heavily on simulations to achieve proficiency while reducing the cost and time requirements of training. Simulations can reduce cost and time compared to use of large-scale exercises at sea, but the assessment of proficiency before, during, and following simulation training is problematic, usually based on observations by human trainers followed by post-training feedback in an "after action review." In this paper we report on studies of the use of computer-based simulations built with the University of Southern California Center for

Cognitive Technology's *iRides* authoring system (Munro, 2004) to assess proficiency in a Navy task taught with simulation-based training: air defense planning, the stationing of Navy assets including ships and airplanes to defend against threats. If reliable and valid simulation-based assessments can be developed, it may be possible to insert them in the simulation-based training to enable real-time assessment of knowledge and skill based on performance of the targeted task. To investigate the feasibility of using such real-time assessment results to enable real-time instructional decisions, the studies also address the use of methods for automating the fusion of assessment results from different sources to support a training assignment decision.

Our research questions were:

- What is the quality of the measures?
- What is the relation between experience and measures of performance?

The first part of this report describes the SWOS task used to focus our assessment system development. We then describe the assessment design process, including the relationship of measures to cognitive demands, the assessments, and the automated scoring methods and representational formats examined in our research. We close with the results of a pilot test conducted with one class attending the Navy's Surface Warfare Officers School, followed by a discussion of lessons learned and next steps.

Method

The task is air defense planning. The student is presented with a situation defined by a location near a coastline, threats posed by enemy airbases with aircraft having certain ranges and carrying certain weapons, and a set of assets—ships and aircraft—that can be used to defend key assets, an aircraft carrier and an "oiler" or supply ship. The student is asked to create an air defense plan by stationing assets in appropriate locations and defining appropriate defensive roles for each asset.

Cognitive Demand

Although it is called a planning task, air defense planning is really a design task, requiring creation of a new organization of components (assets and roles) to satisfy requirements. The design is based on conceptual knowledge of component

capabilities and constraints. In the air defense task, the student must know what critical assets are to be protected (the carrier and the oiler); the locations and capabilities of threats (aircraft and missile launchers and their ranges, the weapons carried by aircraft and their ranges); the defensive capabilities of available ships and airplanes (weapons and their ranges, aircraft ranges, sensor capabilities); basic elements of an air defense design (vital area, surveillance area, threat axis, threat sectors, roles that must be assigned, and assets that can play each role); and tactical details such as speed and heading of assets. To create the air defense plan, or design, the student must use this conceptual knowledge and knowledge of general design constraints (e.g., the oiler should be placed behind the carrier, away from the threat axis, an asset cannot be placed within territorial waters—closer than 12 miles from the coastline of most countries), general procedural and heuristic knowledge (e.g., factor the design into subproblems and watch for interactions between subproblems), and specific knowledge based on assessment of the situation (e.g., geographic features and threat types) that may lead to a solution schema, which may lead to expectancies and a focus of attention on specific cues to confirm expectancies and evaluate the schema.

The cognitive processing required to perform this task is at a high level, involving activity that would be placed in the Create and Evaluate categories of Anderson and Krathwohl (2001) and in Mayer's (1992) Strategic and Schematic categories, and van Merriënboer (1997) would call them examples of strategic knowledge supported by cognitive schemata. The design activity depends on supporting activities that would be classified as Anderson and Krathwohl's Analyze and Apply, and it involves knowledge that can be classified in their Metacognitive, Procedural, and Conceptual knowledge categories. Whatever the classification framework or the terms used to define the activities, it is clear that the task requires high-level, complex cognitive skill based on conceptual knowledge of component capabilities and constraints, and assessments must elicit and measure that skill and knowledge.

The Assessments

The air defense planning skills were assessed using a simulation that attempted to mimic the actual planning task. Students were presented with a series of scenarios presenting situations and a set of assets, and they were asked to develop a plan

(design) for each scenario. The simulation recorded placement of assets and assignment of roles to assets, and scored completed designs against expert designs.

The simulations assess the high-level cognitive skill required to perform the design activities. This cognitive processing requires conceptual knowledge of component capabilities and constraints, knowledge which is not directly measured by the simulations, which presumably measure the results of applying conceptual knowledge, not the knowledge itself. To measure knowledge, we used surveys of student background and knowledge of threat assessment, and a knowledge mapping assessment of knowledge of the threat detection-to-engage process. We also collected self-ratings of air defense knowledge. In addition to providing a measure of conceptual knowledge, the results of these assessments can be compared to the simulation results to provide information on an assessment-criterion relationship, one consideration for validation (Linn & Gronlund, 2000). Table 1 summarizes the relation of cognitive demands to measures and assessment formats used.

Table 1

Cognitive Demand and Associated Toolset Application

Cognitive Demand	Measure	Format
Knowledge of air defense ship and aircraft stationing	Job experience in various Combat Information Center (CIC) roles	Survey
	Level-of-threat detection	Survey
	Detect-to-engage process	Knowledge mapping
Application of air defense stationing knowledge	Air defense ship and air stationing	Computer simulation (iRIDES)

Procedure

A pilot test was conducted to test measures, procedures, and scoring. This test was administered before students received instruction on air defense planning. The purpose of the pretest was threefold: first, to test the administration procedures and software, and gather student and instructor feedback about the tasks; second, to gather preliminary data on our measures for the purpose of refining our measurement approach; and finally, to the extent that the data allowed, to develop a

model to identify students with high prior knowledge of the domain, for the purpose of potential opting out of the course.

Sixty-one Naval officers enrolled in the September 2005 Department Head course participated in this study. Table 2 shows the mean years in service and self-reported air defense knowledge. Table 3 reports the percentage of participants with various qualifications, and Table 4 shows the percentage of participants with various CIC experience.

Table 2

Participant Demographics

Variable	*N*	*M*	*SD*
Years in service	61	11.18	4.89
Self-reported estimate of air defense knowledge[a]	60	1.67	0.84

[a]1 = *little or no knowledge*, 3 = *somewhat knowledgeable*, 5 = *very knowledgeable*.

Table 3

Percentage of Participants Reporting Qualifications (n = 61)

Qualifications	Pct.
CIC	95
SWOS Division Officer	93
SUWC/SSWC	44
ECO	28
TAO	38
AIR/WCO/SWC	36

Table 4

Percentage of Participants Reporting Different Tactical Watch Duties in the CIC (n = 61)

Tactical watch	Pct.
Tactical Action Officer (TAO), CV Combat Direction Center TAO	25
CIC Watch Officer	72
Anti-Air Warfare Coordinator (AAWC/AIR), FFG Weapons Control	20
Tactical Information Coordinator (TIC)	8
Air Resource Element Coordinator (AREC)	7
Identification Supervisor (IDS)	8
Air Control Supervisor (ACS)	7
Air Intercept Controller (AIC	7
Electronic Warfare Supervisor (EWS)	11
EW Console Operator (EWCO)	10
Missile System Supervisor (MSS)	10

We used a single-group, pretest-posttest design. Participants were from an intact class and were administered assessment tasks before and after an instructional period of eight weeks. During the instructional period, the air defense topics were covered as part of a much larger curriculum.

Participants were administered a written survey, then three computer-based tasks. The computer-based tasks were (a) a survey of air defense threat assessment knowledge, (b) an iRIDES simulation-based task requiring students to create an air defense plan, and (c) a knowledge mapping assessment of detect-to-engage process knowledge. Students were administered tasks in two classrooms. Tasks were counterbalanced across classrooms. Table 5 shows the task administration schedule and time allotted for each task.

Table 5

Administration Schedule

Task	Time allotted
Introduction	5
Background information	5
Air defense threat assessment	20
Air defense planning	25
Detect-to-engage knowledge map	30

Measures

Measures produced by the assessment tasks are described in the paragraphs below, in the order shown in Table 5. These descriptions are followed by descriptions of two additional measures: scores on the air defense posttest administered by SWOS as part of the Department Head course, and self-estimates of prior knowledge following the posttest.

Experience in Combat Information Center (CIC). In the survey of background information, participants were asked to report on various aspects of their background (e.g., demographic information, tactical training and schools, qualifications, and air defense experience within the CIC).

Knowledge of threat assessments. Measures of threat detection in littoral and open environments and knowledge radar emitters were adopted from Liebhaber, Kobus, and Feher (2002). The measures for threat detection in littoral and open environments were dropped due to low reliability in both the pretest and posttest (α in the .40s or lower). For the radar emitter survey, several items were added after the pretest, and items contributing to low reliability were dropped from the scale. Reliabilities of the final scale were .82 (pretest, 5 items) and .86 (posttest, 6 items). The final scale is shown in Figure 1.

Radar (ES) Emitters—Please rate, according to your experience, by checking the description that best fits in the blank.

This emitter is _____ associated with a potentially threatening platform.

	don't know	always	often	sometimes	hardly ever	never
13. APG-63	☐	☐	☐	☐	☐	☐
14. APS-124	☐	☐	☐	☐	☐	☐
15. APS-115	☐	☐	☐	☐	☐	☐
16. ARINC-564	☐	☐	☐	☐	☐	☐
17. CAS Search	☐	☐	☐	☐	☐	☐
18. APQ-120	☐	☐	☐	☐	☐	☐
19. Cyrano IV	☐	☐	☐	☐	☐	☐
20. Decca-170	☐	☐	☐	☐	☐	☐
21. Decca-1226	☐	☐	☐	☐	☐	☐
22. Don-2	☐	☐	☐	☐	☐	☐
23. Square Tie	☐	☐	☐	☐	☐	☐
24. Primus-40	☐	☐	☐	☐	☐	☐
25. RDR-1E	☐	☐	☐	☐	☐	☐
26. SPS-10B	☐	☐	☐	☐	☐	☐
27. SPS-55	☐	☐	☐	☐	☐	☐
28. SPS-49	☐	☐	☐	☐	☐	☐
29. Type-992Q	☐	☐	☐	☐	☐	☐
30. Type-1006	☐	☐	☐	☐	☐	☐
31. WM-28	☐	☐	☐	☐	☐	☐

Figure 1. The final radar threat emitters scale, adapted from Liebhaber et al. (2002).

Air defense planning. As noted above, air defense planning involves stationing a set of assets (ships and aircraft) to defend key assets from threats while meeting constraints imposed on basic elements of an air defense plan, including defense areas and sectors, required roles to be played by assets, relative positions of assets playing certain roles, and territorial boundaries. We assess this skill using a simulation developed by USC/CCT in their iRides simulation authoring system. Figure 2 shows the simulation screen with annotations added to describe features and functionality.

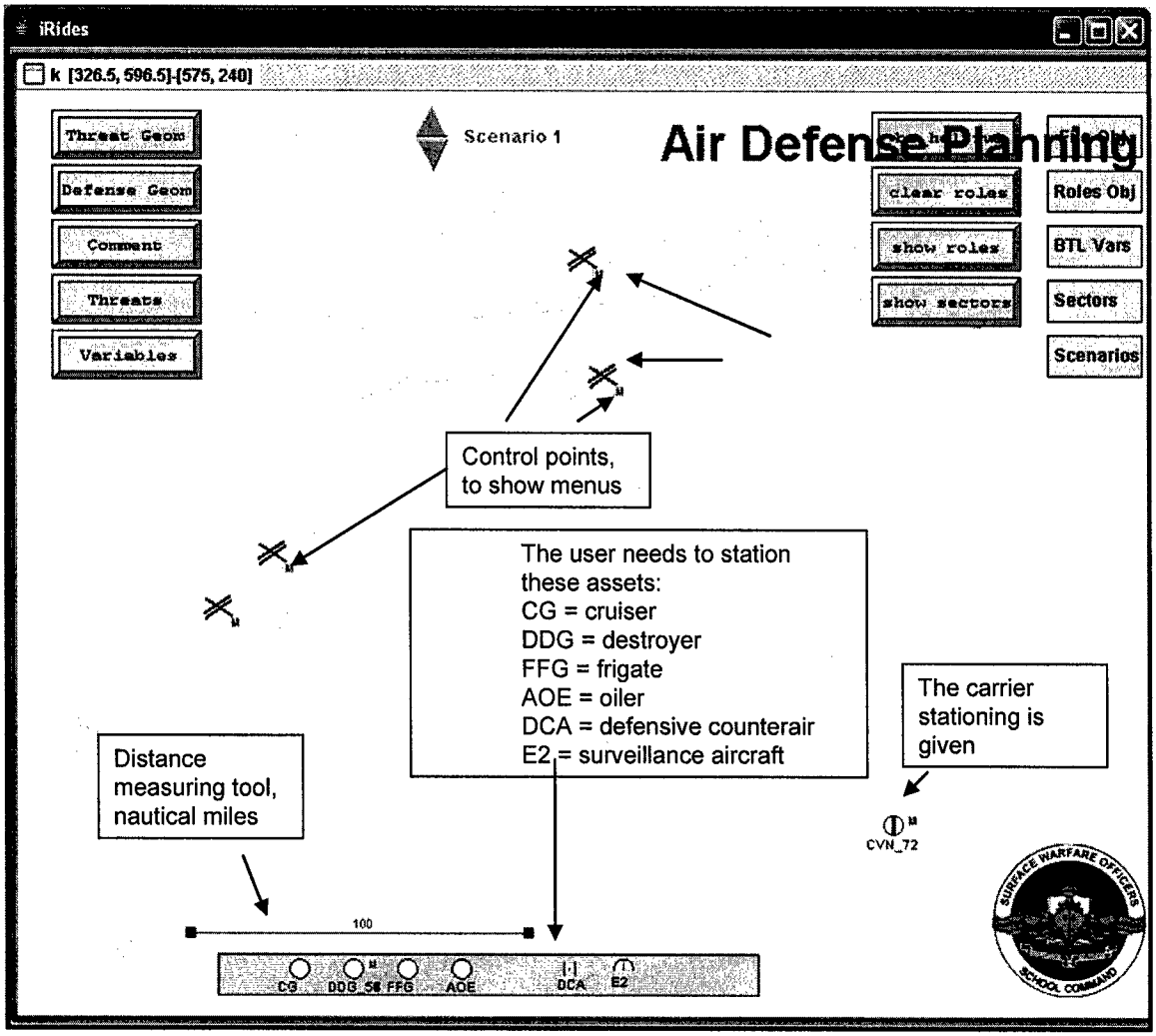

Figure 2. An air defense simulation screen shot with annotations.

The student completes three scenarios, each beginning with a carrier stationed off a coastline with enemy bases. Scenarios differ in the shape of the coastline, the location and capabilities of enemy bases, and the location of the carrier. The student's task is to station the available assets by dragging them from the palette at the bottom of the screen to appropriate locations in order to defend the carrier and the "oiler" or AOE, which also must be placed. A role must be defined for each asset using a popup menu accessible when the icon for an asset is selected. Defensive requirements are influenced by the threat posed by the bases on shore, which can be viewed by selecting the icon for each base.

Air Defense Simulation Scoring

- Role assignment. For each platform, a check is made on whether the student's role assignments are valid. The number of correct role

assignments were added across the three scenarios, even if they only finished one or two, yielding one measure, and the number of incorrect or missing role assignments was the second measure.

- Sector placement. The student's response is compared against the expert's placements, yielding two measures: (a) the number of matches with respect to the correct sector placement, and (b) the number of incorrect or missing placements. Again, both are composites of the placements across one to three scenarios.

- Rules. The third set of measures relate to specific rules about defining the vital, classification, identification, and engagement area (CIEA), and surveillance areas as a user set distance (radius) from the carrier. For each rule, the measure is the count of the number of +1 and -1. The +1 is associated with better (or rule satisfaction) performance, and -1 associated with poorer (or rule violation) performance.

Knowledge of the detect-to-engage process. Knowledge maps were used to measure participants' understanding of the detect-to-engage process. As shown in Table 6, there were 33 concepts and nine relations.

Table 6

Detect-to-Engage Knowledge Map Node and Links

Types of concepts					
Events	Information sources	Purpose	Tools	Actors	Relations
Engagement, ID, Intent, Maneuver	Air Route, Altitude Change, Country of Origin, ECM, IFF, Illuminate (response to), I&W, O/S Caps and Lims, Profile, Query, Speed, Threat Capabilities, VID, Warn (response to warning), Wind, Wings Call	Employ Chaff, Reduce RCS, Unmask Batteries	CIWS, Decoys, Guns, Surface to Air Missile	AIC, Air, DCA, OOD, TAO	conducts, determines, has, orders, part of, recommends, supervises, to, uses

Figure 3 shows a screen shot of an expert's map. The maps were a variation of graphical network representation in that the physical placement of "event" nodes was significant. That is, the plotting area was divided into three bands, where each band corresponded to distance (far, medium, close). Participants were instructed to

place the appropriate "event" node in a band where the event is likely to occur. With the exception of event node placement, the task remained the same (i.e., participants were required to link concepts together with relations) and the physical placement of the other nodes was not significant.

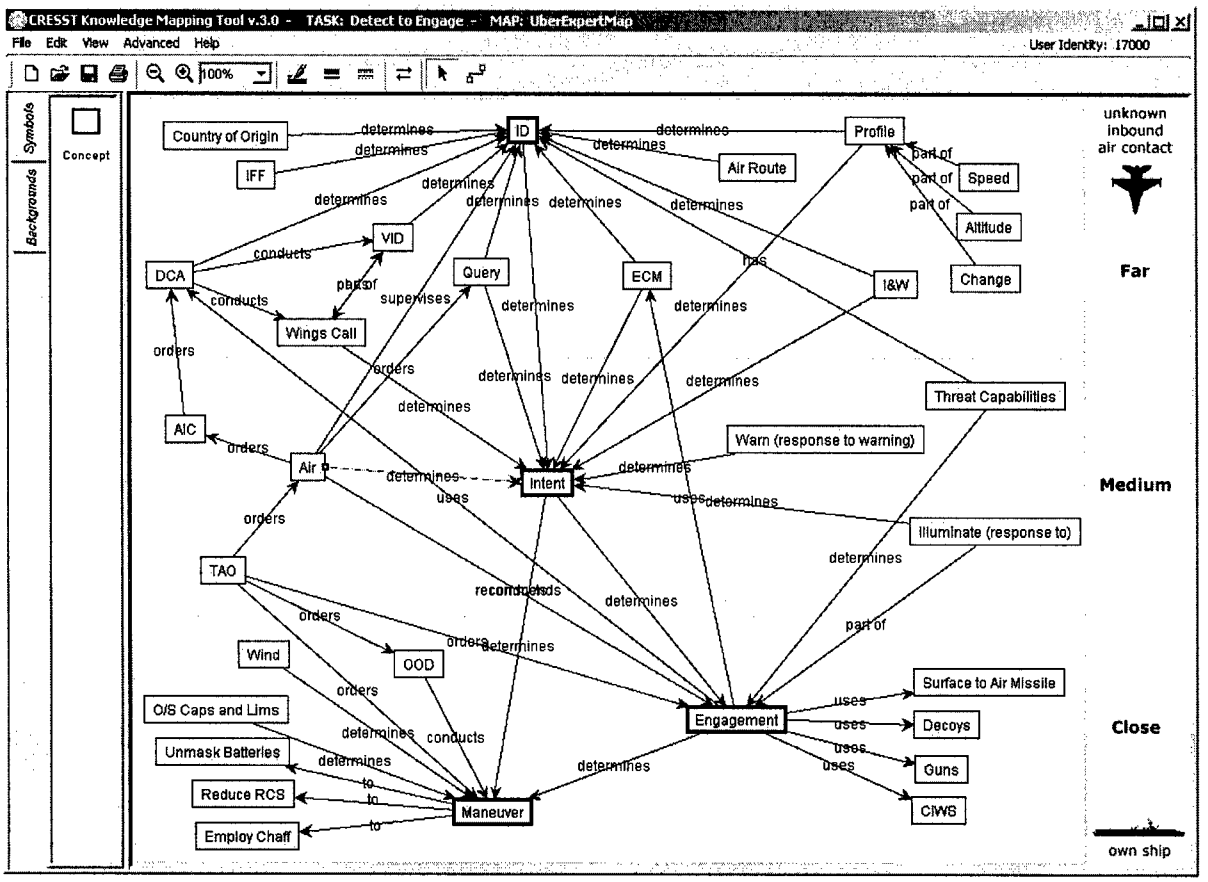

Figure 3. An expert detect-to-engage knowledge map.

Knowledge Map Scoring

- Band placement. The first score was the number of event nodes placed in the appropriate band. The appropriate band for the event *ID* (identification) was the far band, *intent* the medium band, and *maneuver* and *engagement* the close band.

- Comparison to criterion maps. There were two scores based on comparisons to criterion maps—a stringent score based on the original criterion map, and a second score based on participant-generated propositions not identified initially by our experts, but similar to the propositions in the original criterion map.

- Comparison to merged expert's criterion map. The second score was based on comparing participants' knowledge maps to a criterion map. The criterion map was a merging of two experts' maps resulting in 53 unique propositions. The score assigned to a participant map was the number of propositions (node-link-node) in the participant's map that was also in the expert-based criterion map. The maximum possible score was 53.

- Comparison to student-generated "similar propositions" map. The third knowledge map score was computed using a second criterion map. In this case, the second criterion map was based on propositions that participants created in the pretest that were not part of the original criterion map but were similar to propositions in the criterion map. A review of all of the propositions in the student maps showed there was a high frequency for some student-generated propositions that reflected some of the knowledge in the criterion map, but with the thought expressed slightly differently, usually with the same two concepts but connected by a different—though similar—link. We took an approach similar to that of Herl, Baker, and Niemi (1996) where links were sorted into categories, and we grouped links that served similar functions (e.g., *orders, supervises,* and *recommends).* Then, for each proposition in the criterion map, we substituted the synonym links and asked experts to review those new propositions for accuracy. For example, the criterion map has *Air–orders–AIC;* we asked if *Air–supervises–AIC, Air–recommends–AIC,* and *Air–supervises–DCA* could also be considered correct. In addition, there was a subset of participant-generated propositions that conflated two expert map propositions. For example, the expert map has *Speed–part of–Profile,* and *Profile–determines–Intent.* The participant-generated proposition *Speed–determines–Intent* eliminated the intermediate concept *Profile.*

SWOS end-of-unit exams, air defense. We examined end-of-unit scores and scores of items on the end-of-unit exams that were related to air defense. Because there were only three subscale items for air defense, their reliabilities were poor, and the items required lower levels of conceptual knowledge (in our judgment) compared to our tasks, we dropped the subscales from the analyses and used only the total scores.

Self-estimates of prior knowledge (posttest only). A measure was developed to gather participants' perceptions of their level of knowledge of the topics covered related to air defense. Participants were asked to provide ratings related to four questions with respect to each of 18 air defense lessons:

1. How much knowledge did you have prior to taking the lesson?

2. How much new material did you learn from the lesson?
3. Based on what you knew at the beginning of the course, could you have skipped the lesson?
4. Assuming you skipped the lesson, how much impact would it have had on your knowledge?

For each question, internal consistency (Cronbach's alpha) across the 18 items was computed. For air defense, α was .93, .96, .97, and .82 respectively. Given the high reliabilities, a single scale was formed for each topic by combining participants' responses across the four questions. This measure was treated as the main dependent variable for predicting participants' level of pre-instruction knowledge.

Results

Because there was no outcome measure, we used experience as the main proxy for knowledge. *p* values were set to .05, one-tailed. The distributions were in general skewed; therefore, non-parametric correlations (Spearman) were used.

To evaluate the quality of our assessments, we examined how the pattern of correlations among the measures changed between pre- and posttest administrations of the measures, and compared group performance on these measures across pretest and posttest administrations. We expected correlations among measures of similar cognitive demands (e.g., surface knowledge or deep understanding) to be related more than across different cognitive demands. We also expected learning to occur over the instructional period. The change in participants' learning would manifest itself in two ways: (a) a change in the pattern of correlations among the measures, and (b) higher performance on the posttest measures. We also expected a pervasive influence of experience on both types of knowledge, and expected to observe relations between measures of experience and measures of knowledge in general.

Experience

Because we expected experience to be an important predictor of whether participants had the requisite knowledge for each domain, we examined the overall relationship between self-reported experience and the various measures. Results showed that the higher the number of months as Antiair Warfare Coordinator (AAWC) or FFG Weapons Control in a Combat Information Center (CIC), the higher

the participants' overall self-estimated prior knowledge of air defense (r_{sp} = .49, p < .01). This finding is consistent with our expectations of more relevant prior knowledge acquired from on-the-job experience in particular CIC watches. Experience is also related to posttest detect-to-engage (DTE) process knowledge map scores (r_{sp} = .27, p < .01), but, surprisingly, to no other air defense measures.

Pre- vs. Posttest

As predicted, scores increased from pre- to posttest for the air defense simulation (t = 3.76, p < .01) and the radar emitter survey (t = 3.46, p < .01). There was no pre- to posttest difference, however, on the DTE knowledge map scores, and scores actually decreased for the event location knowledge map (t = 3.1, p < .01).

Intercorrelations

Significant correlations were found between DTE knowledge map pre- and posttest scores (r_{sp} = .44, p < .01), DTE pretest and overall air defense simulation pretest scores (r_{sp} = .50, p < .01), and DTE knowledge map posttest scores and radar emitter threat posttest scores (r_{sp} = .36, p < .05). Intercorrelations also indicated that the three air defense simulation scenarios were different. Scenario 1 performance was related to radar emitter threat pretest (r_{sp} = .33, p < .05) and posttest (r_{sp} = .82, p < .01). Scenario 2 performance was related to the pretest overall simulation score (r_{sp} = .27, p < .05), the end-of-unit test score (r_{sp} = .33, p < .01), the radar emitter threat posttest score (r_{sp} = .87, p < .01), and the DTE knowledge map posttest score (r_{sp} = .56, p < .01). Performance on Scenario 3 was related to the posttest radar emitter threat score (r_{sp} = .82, p < .01), the posttest DTE knowledge map score (r_{sp} = .51, p < .01), and the posttest event location knowledge map score (r_{sp} = .65, p < .01).

Discussion

Pilot test results were mixed. As expected, experience had a strong positive effect on performance, but it correlated only with self-ratings of knowledge and the posttest DTE knowledge map scores. Further, although some measures—the air defense simulation and the radar emitter survey—were sensitive to instruction, showing an increase from pre- to posttest scores, many were not, and only a few measures showed the expected intercorrelations.

The purpose of a pilot test is, of course, to identify and revise poor assessments, and these results allowed us to do that. Interviews with pilot test participants also helped us identify problems with the user interfaces of the assessment tools. A subsequent test using revised assessments and improved assessment tool user interfaces produced better results. This test will be described in a subsequent report, but preliminary analyses suggest significant correlations between the air defense simulation scores and experience, self-rated air defense knowledge in most areas, and radar emitter threat scores.

In addition to indicating needed revisions of assessments and the user interfaces of assessment tools, the pilot test results suggest a need to investigate differences among simulation scenarios. The results show that the three scenarios produced different performance. Next steps in this research should include an investigation of the causes of performance differences. We speculate that geography is one important factor, and an important effect of geography is the number of applicable cases in the student's knowledge base. Scenarios 1 and 2 were geographically similar, set in a fairly open area of water along a fairly straight coastline. Examples are Narragansett Bay and the coast of southern California. Scenario 3 was placed in a narrow gulf, with a relatively small area of water enclosed by a coastline on three sides, on which both enemy and friendly countries were located. The small, enclosed area creates special problems in stationing assets, especially when some countries allow flyover and approach within territorial waters and others do not. We believe that this problem is difficult primarily because most students have not encountered them in prior experience and therefore do not have knowledge based on similar cases that can be applied. These are only speculations based on an ad hoc analysis of the scenarios, however, and future research should study the influence of geography and other elements differentiating air defense scenarios.

Future research should also investigate the use of assessment results for performance predictions supporting placement and development decisions. The approach will be similar to that used in prior research on rifle marksmanship (Chung, Delacruz, & Bewley, 2004). Using results from the air defense assessments and self-ratings of air defense experience, we will develop and test Bayesian inference networks to generate the probability that an unobservable variable is in a particular state (e.g., the student understands air defense asset stationing) given observable evidence (student performance on the various measures).

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Chung, G. K. W. K., Delacruz, G. C., & Bewley, W. L. (2004). Performance assessment models and tools for complex tasks. *International Test and Evaluation Association (ITEA) Journal, 25*(1), 47–52.

Herl, H., Baker, E., & Niemi, D. (1996). Construct validation of approach to modeling cognitive structure of US history knowledge. *Journal of Education Research*, 89, 213-230.

Liebhaber, M. J., Kobus, D. A., & Feher, B. A. (2002). *Studies of U.S. Navy cues, information order, and impact of conflicting data* (Tech. Rpt. 1888). SSC San Diego: San Diego, CA.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching, 7th Edition*. Upper Saddle River NJ: Prentice-Hall, Inc.

Mayer, R.E. (1992). *Thinking, problem solving, cognition* (2nd Ed.). New York: W.H. Freeman.

Munro, A. (2004, April). *Authoring simulations that can be integrated with instruction.* Presentation at the AERA 2004 Annual Meeting TICL Symposium I: Advances in Adaptive Authoring Systems in San Diego, CA.

van Merriënboer, J. J. G. (1997). *Training complex cognitive skills.* Englewood Cliffs: Educational Technology Publications.