Experimental Assessment of Accuracy of Automated Knowledge Capture

Susan M. Stevens, J. Chris Forsythe, Robert G. Abbott, and Charles J. Gieseler

Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185, USA
{smsteve,jcforsy,rgabbot,cjgiese}@sandia.gov

Abstract. The U.S. armed services are widely adopting simulation-based training, largely to reduce costs associated with live training. However simulation-based training still requires a high instructor-to-student ratio which is expensive. Intelligent tutoring systems target this need, but they are often associated with high costs for knowledge engineering and implementation. To reduce these costs, we are investigating the use of machine learning to produce models of expert behavior for automated student assessment. A key concern about the expert modeling approach is whether it can provide accurate assessments on complex tasks of real-world interest. This study evaluates of the accuracy of model-based assessments on a complex task. We trained employees at Sandia National Laboratories on a Navy simulator and then compared their simulation performance to the performance of experts using both automated and manual assessment. Results show that automated assessments were comparable to the manual assessments on three metrics.

Keywords: Automated assessment, Naval training systems, simulation-based training, intelligent tutoring systems.

1 Introduction

A significant cost in simulation-based training is the workload on human instructors to monitor student actions and provide corrective feedback. For example, the U.S. Navy trains Naval Flight Officers for the E2-Hawkeye aircraft using a high-fidelity Weapons Systems Trainer (E2 WST). Currently this requires a separate instructor to observe each student within the context of team performance and provide instruction based on observed misunderstandings, inefficient task execution, ineffective or inappropriate actions, etc. Individualized instruction contributes to high training costs. Intelligent tutoring systems target this need, but they are often associated with high costs for knowledge engineering and implementation. New technologies are required that assist instructors in providing individually-relevant instruction.

1.1 Simulation Training

Establishing the validity of automated assessments requires studies in a realistic training environment, rather than just a simple laboratory task. E2 operators are trained and tested on several different simulators ranging from a part-task computer-based

D.D. Schmorrow et al. (Eds.): Augmented Cognition, HCII 2009, LNAI 5638, pp. 212–216, 2009.

training (CBT) system that runs on a single PC, to the high-end E2 WST system which faithfully replicates most aspects of E2 operations (ranging from the physical controls to system fault diagnosis and recovery) and requires a team of instructors of operators to conduct training. For this study we used the E2 Distributed Readiness Trainer (EDRT), a medium-fidelity trainer which presents students with the same mission software used on the E2 aircraft. Multiple instructors are needed to evaluate simulation training and sessions can last hours at a time. Automated assessment of E2 operator performance in these sessions would greatly reduce instructor workload and would increase overall efficiency.

1.2 AEMASE

Sandia National Laboratories has shown the feasibility of automated performance assessment tools such as the Sandia-developed Automated Expert Modeling and Student Evaluation (AEMASE) software. One technique employed by AEMASE is the grading of students performance by comparing their actions to a model of expert behavior. Models of expert behavior are derived by collecting sample data from simulator exercises or other means and then employing machine learning techniques to capture patterns of expert performance. During training, the student behavior is compared to the expert model to identify and target training to individual deficiencies. Another technique utilized by AEMASE is the grading of students performance by comparing their actions to models of good and/or poor student performance. Students with good and bad performance are identified and machine learning techniques are employed to construct models of these two types of performance in the same manner as expert performance. Student performance from other training sessions is then compared to these models to identify and target training to individual deficiencies. Both techniques avoid the costly and time-intensive process of manual knowledge elicitation and expert system implementation (Abbott, 2006).

In a pilot study, AEMASE achieved a high degree of agreement with a human grader (89%) in assessing tactical air engagement scenarios (Abbott, 2006). However, the 68 trials assessed utilized only four subjects under three different initial training scenarios and the range of correct behaviors was quite limited. The current study provides a more rigorous empirical evaluation of the accuracy of these assessments. User modeling, based on behavioral and/or physiological measures, will be a key component of technologies implementing augmented cognition tools for training.

Purpose of Study. Automated assessments, such as AEMASE, would be a helpful tool in assessing E2 operator performance in an EDRT. Using AEMASE, user models can be derived with data generated from students executing scenarios within a simulation trainer or on actual equipment platforms. We trained employees at Sandia National Laboratories on an EDRT and then assessed their simulation performance using both AEMASE and manual assessment.

2 Methods

2.1 Participants

Twelve employees from Sandia National Laboratories volunteered to participate in the experiment. The participants met certain required criteria for the experiment

which reflected the requirements for an entry-level E2 Hawkeye operator. In addition, two former E2 Hawkeye operators participated in the experiment and served as subject matter experts (SME's).

2.2 Materials

Materials included an E2 Deployment Readiness Trainer (EDRT) simulator that was obtained from the Naval Air Systems Command's Manned Flight Simulator organization. The Joint Semi-Automated Forces (JSAF) simulation software was used to create and drive the training and testing scenarios. In addition, the Sandia-developed Automated Expert Modeling and Student Evaluation (AEMASE) software and the Command Distributed Mission Training System (CDMTS) software were used in the analyses of the data.

2.3 Procedure

The participants were recruited via an advertisement and those who responded positively and met the required criteria were included in the study. The participants were scheduled for an initial all-day training session in which a former E2 Hawkeye Naval Flight Officer provided a tutorial on E2 operations emphasizing the basic radar systems task that would be the subject of the experiment. The participants were also asked to sign an informed consent. After the initial training session, the participants were scheduled for seven additional training sessions. The participants were lead through the sessions in the same order. Once they had finished the training sessions, the participants completed two testing sessions. The participants completed the seven training and two testing sessions individually.

Training Sessions. The first five sessions consisted of additional training sessions designed to teach the participants the basic operations of the E2 radar system in depth on the EDRT. For each session, the experimenters first demonstrated the proximate operation(s) on the EDRT and then the participant was asked to perform the operation(s) in scaled down, yet realistic, simulations. Since all five of these sessions were for training purposes, the experimenters were available to answer questions. At the end of each training session, the participants filled out a questionnaire indicating their understanding of the operation(s) on the preceding training session. At the end of the fifth scenario, the participants completed a questionnaire assessing their knowledge of all of the operations learned in the training sessions.

Testing Sessions. The last two sessions were testing sessions in which the participants were assessed on their knowledge of the operations and tactics covered in the five training sessions. The participants completed these more difficult simulations without the help of the experimenters. At the end of each testing session, the participants were asked to complete a questionnaire which queried their confidence of their performance on the preceding testing scenario.

Metrics. Based on guidance from the SMEs, three metrics were developed which were used to grade the participants' performance on the testing sessions. These metrics included fleet protection, labeling of neutral entities and Combat Air Patrol

(CAP) Asset Management. These metrics were used in both the manual and automated assessments.

Fleet Protection. Participants were instructed to prevent non-friendly entities from nearing the carrier group. The amount of time the non-friendly entities spent too close to the carrier group was assessed.

Labeling neutral entities. Participants were instructed to promptly and appropriately label any neutral entity that appeared on the radar scope. The latency with which the participants took to label these entities was assessed.

Combat Air Patrol (CAP) Assessment. Participants were instructed to effectively manage their air assets as the battle space evolved during the scenario. This included reordering CAP stations so that the airspace would not be violated.

Manual Assessment. Two trained experimenters independently reviewed video recordings of each of the testing scenarios for all participants. The experimenters graded the participants' performance on the three metrics for the two testing scenarios. For each metric, the two experiments specified at least one instance of good and one instance poor student performance. These instances formed subsets of manual assessment data that was used in training the AEMASE system.

Automated Assessment. The participant performance on the two testing scenarios was assessed by AEMASE. AEMASE used the good and poor instances indentified by the two experimenters as base examples from which to assess participant performance.

3 Results

The manual assessments and the automated assessments were compared for each of the three metrics.

Fleet Protection

Manual assessment was based on the amount of time the non-friendly fighters spent too close to the carrier group. The interrater reliability between the two experimenters was 99%. The automated assessment used a proxy measure, which consisted of the distance between the carrier group and the closest non-friendly asset. The results indicate a 100% agreement between the automated and manual assessments in terms of identification of unsatisfactory student performance (i.e., those students whose non-friendly assets got closest to the carrier group).

Labeling of entities

Manual assessment was based on reviewing the timestamped recording of when the neutral entities were labeled. The interrater reliability between the two experimenters was 94%. The automated assessment was based on the analysis of network messages

from the mission computer. The results indicate a 95% agreement between the automated and manual assessment for correct labeling of the neutral entities.

CAP Station Rotation

Manual assessment was based on the time and accuracy with which the CAP stations were reordered. The interrater reliability between the two experimenters was 99%. The automated assessment was based on post-hoc analysis of radio communications. Results indicate an 83% agreement between the automated and manual assessment.

4 Discussion

AEMASE surpassed target performance criteria with agreement of up to 100% with the manual assessment. Even with an undeniably difficult metric that was based on radio communication (the CAP station rotation metric), agreement between AEMASE and manual assessment was an impressive 83%.

Acknowledgements

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was performed through a contract award from the Office of Naval Research.

Reference

1. Abbott, R.G.: Automated expert modeling for automated student evaluation. Intelligent Tutoring Systems, 1–10 (2006)