Automatic Assessment of Complex Task Performance in Games and Simulations

Markus R. Iseli, Alan D. Koenig
University of California Los Angeles
Los Angeles, CA
iseli@cse.ucla.edu, koenig@cse.ucla.edu

John Lee, Richard Wainess
University of California Los Angeles
Los Angeles, CA
johnjn@ucla.edu, wainess@cse.ucla.edu

ABSTRACT

Assessment of complex task performance is crucial to evaluating personnel in critical job functions such as Navy damage control operations aboard ships. Games and simulations can be instrumental in this process, as they can present a broad range of complex scenarios without involving harm to people or property. However, *automatic* performance assessment of complex tasks is challenging, because it involves the modeling and understanding of how experts think when presented with a series of observed in-game actions.

Our previous research was focused on developing a conceptual framework for assessing complex behaviors in non-linear, 3-D computer-based simulation environments. Building on this research, the focus of this paper is on automatic complex task scoring of decision making ability critical to Navy damage control operations. We are using our existing 3-D simulation of the interior of a naval ship (Koenig et al., 2009) which includes both fire-fighting and flooding damage control scenarios. When assessing performance, human expert scoring can be limiting, as it depends on subjective observations of in-game player's performance which in turn is used to interpret their mastery of key associated cognitive constructs.

We introduce a computational framework that incorporates the automatic performance assessment of complex tasks or action sequences as well as the modeling of real-world, simulated, or cognitive processes by modeling player actions, simulation states and events, conditional simulation state transitions, and cognitive construct dependencies using a dynamic Bayesian network. This novel approach combines a state-space model along with a probabilistic framework of Bayesian statistics, which allows us to draw probabilistic inferences about a player's decision making abilities. Through this process, a comparison of human expert scoring and dynamic Bayesian network scoring is presented.

ABOUT THE AUTHORS

Dr. Markus R. Iseli is a Senior Research Associate at UCLA/CRESST with a focus on integration and application of artificial intelligence algorithms for technology-based learning and assessment systems. Prior to working at CRESST, he was a lecturer for digital speech processing at UCLA. He has 10 years of industrial expertise as a technology consultant and hardware and software engineer. Dr. Iseli holds a PhD and MS degree from UCLA and a MS degree from ETH Zurich, Switzerland, all in Electrical Engineering. His specialization is in digital signal processing, speech and image analysis, and pattern matching.

Dr. Alan Koenig is a Senior Research Associate at UCLA/CRESST where he specializes in the application of innovative uses of technology for delivering and assessing instruction. His research focuses on the design and implementation of computer-based games and simulations designed for classroom and/or military training environments. His current work centers on the development of automated assessment systems for high fidelity games and simulations used in the military. Prior to joining CRESST, Dr. Koenig spent 10 years working in the technology sector as both a software developer and mechanical design engineer. Dr. Koenig holds a PhD in Educational Technology from Arizona State University, a BS in Mechanical Engineering from the University of Hartford, and a BA in Economics from the University of Connecticut.

Dr. John J. Lee is a Senior Research Associate at 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 ships 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.

Dr. Richard Wainess, prior to joining CRESST, was a senior lecturer in the University of Southern California's Information Technology Program. His research interests center on the use of games and simulations for training and assessment of adult learners. Richard's most recent work is focused primarily on assessment of problem solving and decision making using computer-based interactive tools. He has authored and co-authored numerous reports, articles, and book chapters and has presented at many conferences on the topic of games and simulations for learning, with a particular emphasis on instructional methods, cognitive load theory, and learning outcomes. His teaching and industry experiences include: digital media design and management; 3D modeling, animation, and visual effects; interactive multimedia production; video writing, producing and directing; and video game design and development.

Automatic Assessment of Complex Task Performance in Games and Simulations

Markus R. Iseli, Alan D. Koenig
University of California Los Angeles
Los Angeles, CA
iseli@cse.ucla.edu, koenig@cse.ucla.edu

John Lee, Richard Wainess
University of California Los Angeles
Los Angeles, CA
johnjn@ucla.edu, wainess@cse.ucla.edu

Assessment of complex task performance is crucial to evaluating personnel in critical job functions such as Navy damage control operations aboard ships. Games and simulations can be instrumental in this process, as they can present a broad range of complex scenarios without involving harm to people or property. When assessing performance, human expert scoring can be limiting, as it depends on subjective observations of in-game player's performance which in turn is used to interpret their mastery of key associated cognitive constructs. On the other hand, *automatic* performance assessment of complex tasks presents its own challenges, because it involves the modeling and understanding of how experts think when presented with a series of observed in-game actions.

Previous research used Bayesian networks to *model cognitive demands and to score performance assessments*. In Chung et al. (2003), performance assessments were tied to instruction using Bayesian networks in the domain of rifle marksmanship. Construction of the Bayesian networks was done using expert knowledge about the domain structure. In the evidence-centered assessment design (ECD) framework, Mislevy et al. (2004) introduced (naïve) Bayesian networks for probability-based reasoning to accumulate evidence of task performances in terms of beliefs about unobservable variables that characterize knowledge, skills, and/or abilities of students. Baker et al. (2008) discussed the design and validation of technology-based performance assessments. They listed expert-based scoring and domain-modeling methods as possible scoring techniques and mentioned the use of Bayesian networks to model student understanding by linking student task performance to latent knowledge and skill states. Almond et al. (2009) described the use of static Bayesian networks for the assessment of proficiency variables in a classroom. Their Bayesian network represents a proficiency model where the nodes are a collection of latent variables and where the students' individual assessment results are entered to yield a total proficiency score for a group of students.
The following publications included dynamic Bayesian networks (DBNs) to *model simulation* or *real-world processes*. Poropudas and Virtanen (2007) used a DBN to model an air combat simulation. They presented a method for analyzing the evolution of discrete events and for learning the network structure and probability tables from simulation data. In neuroimaging (Rajapakse et al., 2007), the data from a functional magnetic resonance imaging (fMRI) scan of brain regions is entered into a DBN to learn the structure of effective brain connectivity between brain regions.

Based on the *conceptual* framework presented in Koenig et al. (2009), this study presents a *computational* framework for automatic performance assessment of complex tasks that allows the combination of *models for cognitive, simulation, and real-world processes* to be united into one DBN. This allows the performance assessment of complex tasks or action sequences as well as the modeling and inference-making of real-world, simulated, or cognitive processes. A description of the computational framework and its procedures for automatic scoring of complex task performance in games and simulations is provided.

THE STUDY

This study presents a proof of concept showing how well expert scoring of complex tasks can be modeled by using a novel computational framework that is represented by a DBN.

In Figure 1, an overview of this study is given. Subject matter experts (SMEs) provide information about how to score player actions in the simulation. This information is then automatically transferred to conditional probability tables of a DBN. In addition, information about the processes in the simulation, as well as dependencies of other processes (real-world, cognitive), help define the state-space topology of the DBN. Once the DBN is constructed, player actions in the simulation are scored by SMEs and by the DBN, yielding expert scores that are compared to DBN scores.

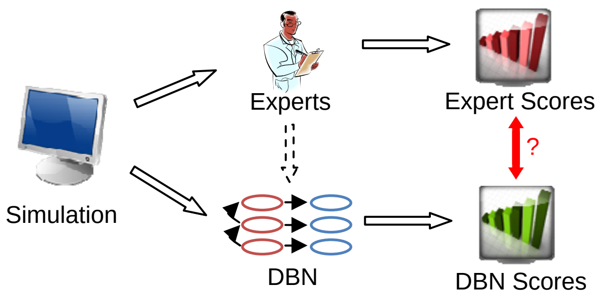

Figure 1 Overview of this study. DBN = dynamic Bayesian network.

METHODS

Our automatic performance assessment system incorporates two parts: (a) a *knowledge-base* that stores SMEs' knowledge, and (b) a state-space model that defines the states of the simulation and their transition over time, given player actions, and game events. Compared to an expert system that is based on an SME knowledge base, our system is capable of adding state-space models of a real-world, simulated, or cognitive processes. It will be shown below that both, SME knowledge-base and state-space models, can be integrated into a single DBN.

The Knowledge-Base

In expert systems, knowledge can be represented as logical statements with associated certainty factors. To use an example from our simulation, the logical statement

> "*If* a player does action A_1 *and then* action A_2 in situation S of the simulation, *then* the player shows a certain knowledge/skill/ability K with a certainty factor of Q%",

shows the SME's reasoning when observing a player's sequence of actions in a given state of simulation and the SME's confidence in the inference of K drawn from the observation. For our purpose of scoring decision making ability, we reformulate above example to:

> "*If* a player does action A_1 *and then* action A_2 in situation S of the simulation, *then* the player shows a decision making ability of Q",

where Q is a value between 0 and 1 using the scoring rubric in Table 1.

Table 1 Scoring Rubric

Score	Description	Q
Optimal	The best action possible	1.0
Good	A good action, but an obvious better one exists	0.85
Adequate	The action correctly addresses the situation, but many better choices exist	0.65
Neutral	The action is unrelated to the situation	0.5
Bad	The action is a bad choice, and has the potential for doing more harm than good	0.0

In order to reduce inter-rater variability the authors formed a panel of "simulation damage control experts" – as opposed to real-life damage control experts - and agreed on the basic rules and scoring rubrics of damage control in our simulation, trying to match the procedures in accordance with Navy doctrine. Our simulation contained four fire situations and four flooding situations: Galley Grease Fire, Storage Room Alpha Fire, Communication Room Electrical Fire, Berthing Area Alpha Fire, Bathroom Fire Main Leak, Bathroom Flood, AFFF Pump Station Leak, and Jet Fuel Pipe Leak. For each situation in our simulation, SMEs created a scoring criteria table that lists all the possible player actions and simulation events in that situation and the necessary conditions on the states of the simulation to determine a score for decision making ability. Table 2 lists the scoring criteria for a fire and a flooding situation. It can be seen that the scores for attacking a burning fire depend on the extinguishing agents used; in this case Aqueous Film-Forming Foam (AFFF), Carbon Dioxide (CO2), "Purple-K Powder" (PKP), and the sprinkler system with Aqueous Potassium Chlorate (APC). The simulation event "re-flash" always indicates that either fire or flood were not correctly overhauled and therefore re-ignited or re-flooded. Scoring criteria for a total of 37 player actions and 8 simulation events were entered into the scoring criteria table.

Table 2 Excerpt from the scoring criteria table for the two situations Galley Grease Fire and AFFF Leak. Note: AFFF = Aqueous Film-Forming Foam, CO2 = carbon dioxide, PKP = "Purple K powder," APC = Aqueous Potassium Chlorate.

Situation: Galley Grease Fire		**Scores**				
		Optimal	Good	Adequate	Neutral	Bad
Actions and Events	Spray AFFF	Fire burning				Fire smoking
	Spray CO2		Fire burning			Fire smoking
	Spray PKP			Fire burning		Fire smoking
	Activate APC	Fire burning				Fire not burning
	De-smoke	Fire smoking				Fire burning
	Event: Re-flash					Always

Situation: AFFF Leak		**Scores**				
		Optimal	Good	Adequate	Neutral	Bad
A.& E.	Patch Leak	Always				
	Overhaul Leak	Always				
	Event: Re-flash					Always

The State-Space Model

The conditions in the scoring criteria table (e.g. "Fire burning") can be represented by a *logical statement* that contains references to object states of the same or of any other situation. For example, patching a leak in situation one (S^1) might be optimal only if the fire in situation two (S^2) has been extinguished and the valve in situation three (S^3) has been turned off. Situations can represent any set of physical compartments on the ship, logical entities, categories, or simulation states used for scoring.

The scoring of player action sequences can be done using (simulation) states to keep track of previous actions. This approach directly leads to the use of state-space models, where the simulation states record previous actions and the performance score of the current action is conditioned on previous simulation states. This approach works well with observable data, but for missing, noisy, or unobservable (latent) data, a probabilistic framework has to be introduced. Dynamic Bayesian networks do exactly this: They represent state-space models using a probabilistic framework.

Dynamic Bayesian Networks

Dynamic Bayesian networks extend Bayesian networks by modeling dynamic systems as opposed to static systems. Dynamic Bayesian networks are versatile representations of state-space models (Murphy, 2002) and can graphically model probabilistic time-dependencies between variables. In the graphical representation as a network, each node represents a variable and each directed link (arrow) represents a dependency between nodes (i.e. node A $\rightarrow$ node B means that variable B is dependent on variable A). By being able to model discrete-time or continuous-time processes, including inputs (e.g. player actions), outputs (observations, simulation events), states (latent and observed), and state transitions of the processes, DBNs can learn both parameters and network structure and can infer or predict unobserved outcomes. There are three approaches to find the structure and probability tables of a DBN: (a) using expert knowledge, (b) using observation data, and (c) a combination of both. In this paper, we will use expert knowledge to determine DBN structure.

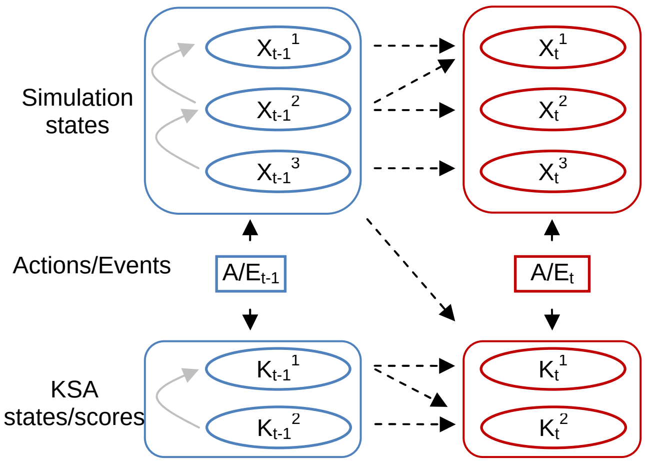

Figure 2 Dynamic Bayesian Network representing dependencies of simulation and knowledge states given an action or event at time *t*. KSA = Knowledge, skill, or ability.

Figure 2 depicts the conceptual overview of the DBN used in our framework. It shows two time slices, at time *t*-1 and time *t* with corresponding actions and states. Arrows in the figure indicate dependencies. Arrows across time slices are dashed, whereas arrows within a time slice are solid. Because our simulation deals with discrete actions and events, the index *t* is increased every time a new action or event happens. In this particular DBN, simulation states, X, are observable, whereas knowledge states, K, are not (i.e. X is an observable variable and K is a latent variable).

Knowledge about the model of the simulation program is stored in the conditional probability tables (CPTs) of the simulation states, where the current (index *t*) simulation state is dependent on previous states and the current action. An example logical statement that represents such a state transition is: if X_{t-1} = "Fire burning" and A_t = "Spray AFFF", then X_t = "Fire smoking."

The scoring rules elicited from the SMEs are stored in the CPTs of the KSA score states and are logical statements like this: if K_{t-1} = "bad" and X_{t-1} = "Fire burning" and A_t = "Spray AFFF", then K_t = "adequate." This means that the current decision making ability score is dependent on previous scores, previous simulation states, and the current action. More dependencies and states can be added. For example, a new state representing the overall fire fighting score and having all states containing fire fighting scores as children could be added.
In this study, for simplicity, we did not assume any dependencies between K_t and K_{t-1} nor between states of the same time slice.

Figure 3 shows an excerpt of our actual DBN designed with GeNIe/SMILE (Version 2.0). It shows the state transitions of some of the fire states going from "burning" to "smoking", to "out." The nodes Node3 to Node6 correspond to the Action/Events (A/E) nodes in Figure 2 and provide the relevant actions and events to state and score nodes.

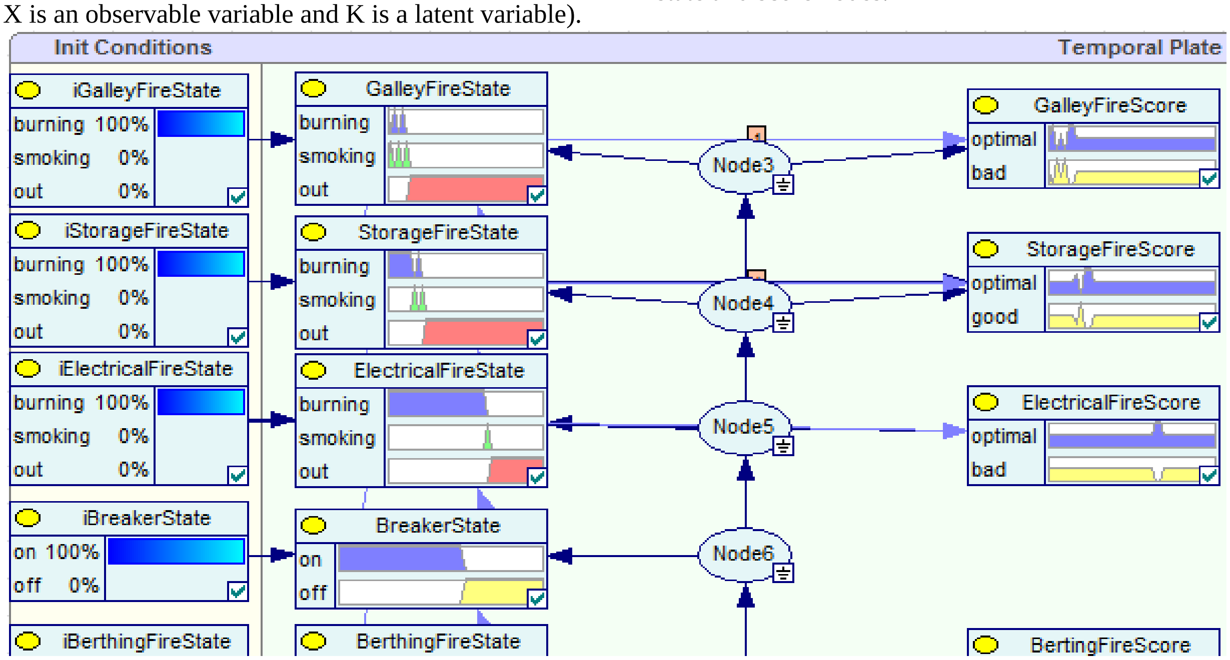

Figure 3 Excerpt from DBN used in this paper. Node3 to Node6 provide actions and events that are relevant for each simulation state or score.

DATA COLLECTION

Participants were recruited from a university as part of an introductory psychology course and participation counted as laboratory credit for their course. The participants were informed of the voluntary nature of the study and that they were able to stop at any point, especially if the participants experienced any dizziness that may have resulted from movement in the 3D game environment.

Simulation data from 30 (9 male, 21 female) participants was collected and analyzed. Of the 30, 56.7% have never played video games, 33.3% play 1-2 hours per week, 6.7% play 3-6 hours per week, and 3.3% play more than 6 hours per week. Fifty percent of the participants said that they were very comfortable using computers, whereas 13.3% stated that they were very uncomfortable.

In order to guarantee well-balanced levels of prior knowledge, participants were randomly assigned to receive one out of four groups of instruction: (a) fire fighting and flooding instruction, (b) fire fighting instruction only, (c) flooding instruction only, and (d) no instruction. Before starting the simulation, they entered a simulation tutorial where they were taught the game mechanics like moving around, opening doors, picking up and dropping equipment. Playing the simulation, participants were asked to discover as many of the eight situations as possible and to address the ones that required some actions. Once done with the simulation, participants filled out a demographic/usability questionnaire in an online format.

The simulation environment used in this study was produced with the Unity 3D game engine. The simulation consisted of a first person perspective 3D environment in which the player could enter different compartments and interact with different objects aboard a Navy ship. This environment allowed for the capture of all player actions and simulation events in real time, which were then fed into the DBN for automatic scoring. For expert scoring, this information was provided in human-readable format to the SMEs for expert scoring.

RESULTS

The goal of this study was to validate the use of automated DBN's in the evaluation of complex performances. To do this, scores were calculated for each player with both human raters (Human) and using the DBN. The human scoring was based on pre-existing Navy doctrine that expert human raters use to evaluate human performance. The DBN scoring was derived from this same criteria and represented using conditional probability tables. Scores ranged from 0 (no player mastery) to 1 (full player mastery; see Table 1).

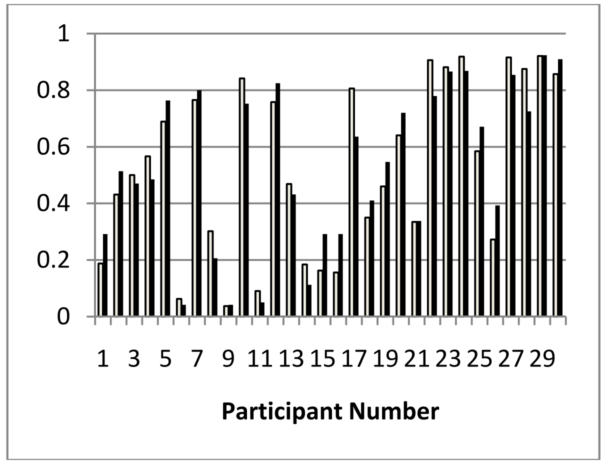

Figure 4 Overall Decision making ability scores: Human (white bars) versus DBN (black bars) scores (Pearson correlation coefficient, r = 0.98)

A total of more than 600 relevant player actions were recorded and scored, resulting in action sequences of about 20 actions for each participant. Aggregates of these scores were calculated for each player and the results are shown in Figure 4 through
Figure 6. Figure 4 shows the players' decision making ability for damage control overall (combined fire fighting decision making and flooding decision making).

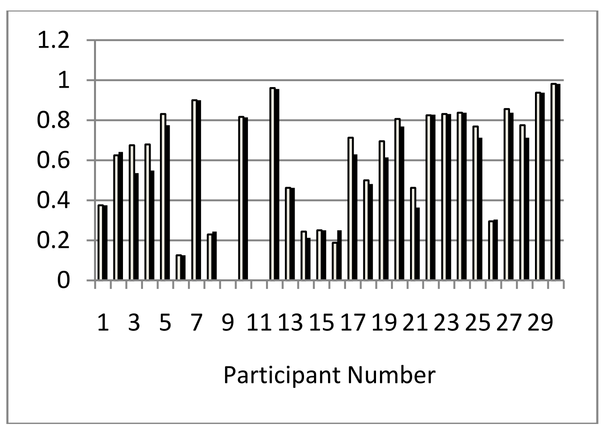

Figure 5 Fire fighting damage control decision making scores: Human (white) versus DBN (black) (r = 0.99)

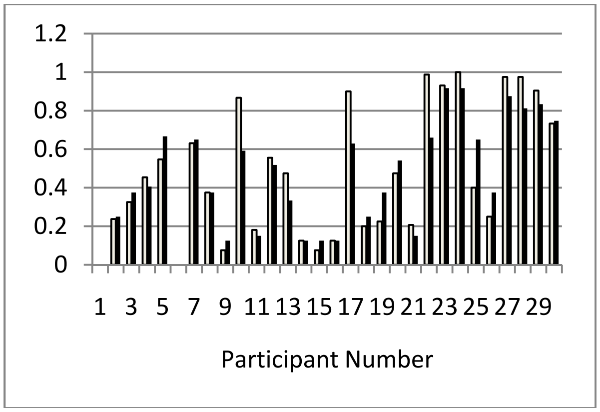

Figure 6 Flooding damage control decision making scores: Human (white) versus DBN (black) (r = 0.97)

Figures 5 and 6 disaggregate the scores by fire fighting and flooding, respectively. As can be seen in the graphs, the human scoring and DBN scoring were very highly correlated with Pearson moment correlation coefficients r=0.98, 0.99, and 0.97, respectively).

In flooding damage control situations, the simulation engine used a leak recurrence time that was too short and unrealistic. In contrast, the SME scoring panel weighed flood recurrences less negatively and thus their scores were generally higher than the DBN scores.

In essence, the discrepancies between the human and DBN scoring were a result of the human scoring being more holistic, tending to focus more on overall performance rather than discrete actions. For example, if a player opened and closed a pipe valve multiple times, the human scoring was more concerned with whether the valve was ultimately left open or closed, whereas the DBN scoring incremented or decremented their score based on each individual action in the order it was done.

In order to calculate inter-rater agreement between human and DBN scores using Cohen's Kappa, the aggregates overall scores from Figure 4 were rounded to the nearest integer. The resulting agreement table is shown in Table 3, where Human-DBN rating agreement counts show on the diagonal and disagreement counts are shown off-diagonal; it can be seen that 23 (diagonal 4+8+2+9) out of 30 participants were rated the same, yielding a rater agreement between Human and DBN of 77% with κ = 0.674.

Table 3 Observed counts of inter-rater agreement on overall decision making ability: Human ratings versus DBN ratings. (κ = 0.674, agreement is 77%)

		DBN rating				
		bad	neutral	adequate	good	Total
Human rating	bad	4	3	0	0	7
	neutral	1	8	0	0	9
	adequate	0	0	2	1	3
	good	0	0	2	9	11
Total		5	11	4	10	30

SUMMARY AND DISCUSSION

The purpose of this study was to validate DBNs for use in the automated scoring of complex tasks. To that end we chose a bounded domain of damage control operations aboard Navy ships consisting of fire fighting and flooding. We worked with Navy SME's to elicit evaluation criteria and used this information to develop our DBN. To validate the DBN, we compared the DBN scores with those from expert human raters.

Overall, there was a high correlation between the two scoring methods. However, the human scored approach tended to be more forgiving on individual constituent actions and was more concerned about holistic outcomes, whereas the DBN was not making these comparisons due to an incomplete holistic representation of expert knowledge in the DBN. The implication of this is that DBNs require a significant level of effort in converting implicit expert knowledge into explicit representations in the DBNs. This in turn might translate into long DBN development lead times.

Despite the high correlations observed, this domain was narrowly bounded and the tasks were specific and well defined. However, there are many cases where the evaluation of human performance involves domains and settings that are much more broad and complex. In those cases, having high correlation between expert raters and a DBN may prove more difficult. Further research is needed to find ways to more efficiently elicit knowledge from experts to be incorporated into DBN's. This would help to make utilization of automated scoring more practical for everyday situations.

The use of the computational framework using a DBN presented in this paper can help reduce or eliminate the need for human raters and decrease the time to score. This has the benefit of potentially reducing costs. In addition, it can facilitate the efficient aggregation, standardization, and reporting of the scores. For these reasons, we encourage continued research in the use of DBN's, especially for military-related evaluations.

We would like to triangulate our results further by using other data collection methods, including non-invasive computer-based eye tracking, after action interviews, and a concept mapping technique called the Cognitive Process Mapper (Wainess, 2008), which enables a student to demonstrate their knowledge of construct relationships in a domain.

ACKNOWLEDGEMENTS

We would like to thank Dr. Robert Atkinson at Arizona State University and his professional staff of laboratory assistants for their support. We would especially like to thank Stacey Joseph at ASU and Matthew Zhang at CRESST/UCLA.

The work reported herein was supported by the Office of Naval Research. However, the findings and opinions expressed in this report do not reflect the positions or policies of the Office of Naval Research.

REFERENCES

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J. -D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning*. Special Section on Bayesian Modelling, 50(3), 450–460.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2008). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. J. G. v. Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 595–604). New York, NY: Erlbaum.

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. Proc. I/ITSEC, 25, 1811–1822.

GeNIe/SMILE (Version 2.0) [Computer software]. Pittsburgh, PA: Decision Systems Laboratory, University of Pittsburgh. Retrieved from http://genie.sis.pitt.edu/.

Koenig, A. D., Lee, J. J., Iseli, M. R., & Wainess, R. A. (2009, November). A conceptual framework for assessing performance in games and simulations. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*, Orlando, FL.

Mislevy, K. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centered design (CSE Tech. Rep. No. 632). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Murphy, K.P. (2002). *Dynamic Bayesian networks: Representation, inference and learning* (Unpublished doctoral dissertation). University of California, Berkeley, Berkeley, CA.

Poropudas, J., & Virtanen, K. (2007). Analyzing air combat simulation results with dynamic Bayesian networks. In *Proceedings of the 2007 Winter Simulation Conference*, pp. 1370–1377. Washington, DC: Institute of Electrical and Electronics Engineers.

Rajapakse, J. C., & Zhou, J. (2007). Learning effective brain connectivity with dynamic Bayesian networks. *Neuroimage*, 37, 749–760.

Wainess, R. (2008, March). *Development and validation of a cognitive process mapper*. Paper presented at the 2008 annual meeting of the American Educational Research Association, New York, NY.