CRESST REPORT 771

Alan D. Koenig

John J. Lee

Markus Iseli

Richard Wainess

A CONCEPTUAL FRAMEWORK FOR ASSESSING PERFORMANCE IN GAMES AND SIMULATIONS

JULY, 2010

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

A Conceptual Framework for Assessing Performance in Games and Simulations

CRESST Report 771

Alan D. Koenig, John J. Lee, Markus Iseli, & Richard Wainess
CRESST/University of California, Los Angeles

July, 2010

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
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by the Office of Naval Research, grant #N0001408C0563, as administered by the Office of Naval Research (ONR), U.S. Department of Defense.

The findings and opinions expressed herein are those of the author(s) and do not necessarily reflect the positions or policies of the Office of Naval Research, or the U. S. Department of Defense.

To cite from this report, please use the following as your APA reference: Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation.* (CRESST Report 771). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

A CONCEPTUAL FRAMEWORK FOR ASSESSING PERFORMANCE IN GAMES AND SIMULATION

Alan D. Koenig, John J. Lee, Markus Iseli, and Richard Wainess
CRESST/University of California, Los Angeles

Abstract

The military's need for high-fidelity games and simulations is substantial, as these environments can be valuable for demonstration of essential knowledge, skills, and abilities required in complex tasks. However assessing performance in these settings can be difficult—particularly in non-linear simulations where more than one pathway to success or failure may exist. The challenge lies not in capturing the raw data arising from game-play, but in interpreting what a player's actions and decisions mean in the broader context of cognitive readiness for a particular job function or task.

The aim of our current research is to develop a conceptual framework for assessing complex behaviors in non-linear, 3-D computer-based simulation environments. Central to this framework is the incorporation of both a domain ontology (which depicts the key constructs and relationships that comprise the domain being simulated), and one or more Bayesian networks (which catalog the probabilities of various sequences of actions related to the constructs in the ontology). For the current research, the domain is damage control related to fire-fighting onboard naval ships, and the two key constructs being assessed are situation awareness and decision-making.

A 3-D, computer-based simulation depicting the interior of a naval ship has been developed. Assuming the role of a damage control investigator, the player is tasked with identifying, addressing, and reporting on a variety of potential, imminent, and existing fires and fire hazards. Using a dynamic Bayesian network, all actions and decisions related to situation awareness, communications, and decision-making are evaluated and recorded in real time, and are used for both formative and summative assessments of performance. Using this conceptual framework, our goal is to provide a generic model of assessment that can be incorporated into both new and pre-existing computer-based simulations that depict cognitively complex scenarios.

Introduction

Endsley (1988; 2000) defines situation awareness 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 the 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. However, cognitive errors in perception can lead to poor decisions, even though the decision itself could well be the correct action for the perception; it is the perception that was incorrect.

To take a query-driven format, the different levels relate to the following questions:

1. **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?
2. **Level 2**: Do you know why the relevant cues are important? Which are not and why? What patterns do you see?
3. **Level 3**: What are you expecting to happen?

Sailors must also be able to think on their feet especially if they are to take initiative when normal standard operating procedures (SOPs) cannot be followed and/or when communications break down or are not possible. Based on Pascual and Henderson (1997), we define decision-making in relation to command and control reducing the number of working practices from 22 to the following 6:

1. Adherence to Standard Operating Procedures (SOPs)
2. Gather additional information
3. Priority/ Risk assessment
4. Task plan/Courses of Action (COAs)
5. Delegation of tasks
6. Monitoring outcome(s) of the COAs

However, both situation awareness and decision-making—as we have defined them here—are difficult constructs to measure directly in a game or simulation. Indeed, the actual data collected from a game typically involves responses to simple triggers that arise in the game in which the player selects objects, allocates resources, interacts with non-player characters, etc. In order to link these basal actions to the higher order constructs of situation awareness and decision-making, we need to devise a conceptual structure that relates all the possible observable, lower level player actions to these more abstract constructs. To do this, we have devised a multi-step process for assessing complex performance in games and simulations. This process is foundationally built off the evidence-centered approach to assessment design (Mislevy, Steinberg, & Almond, 2003), and involves the development of a domain ontology, the construction of a Bayesian network, and the incorporation of various analytic and reporting tools. Outlined below is an overview of the process, along with a description of an initial validation study that was conducted to evaluate the process using the damage control fire-fighting domain.

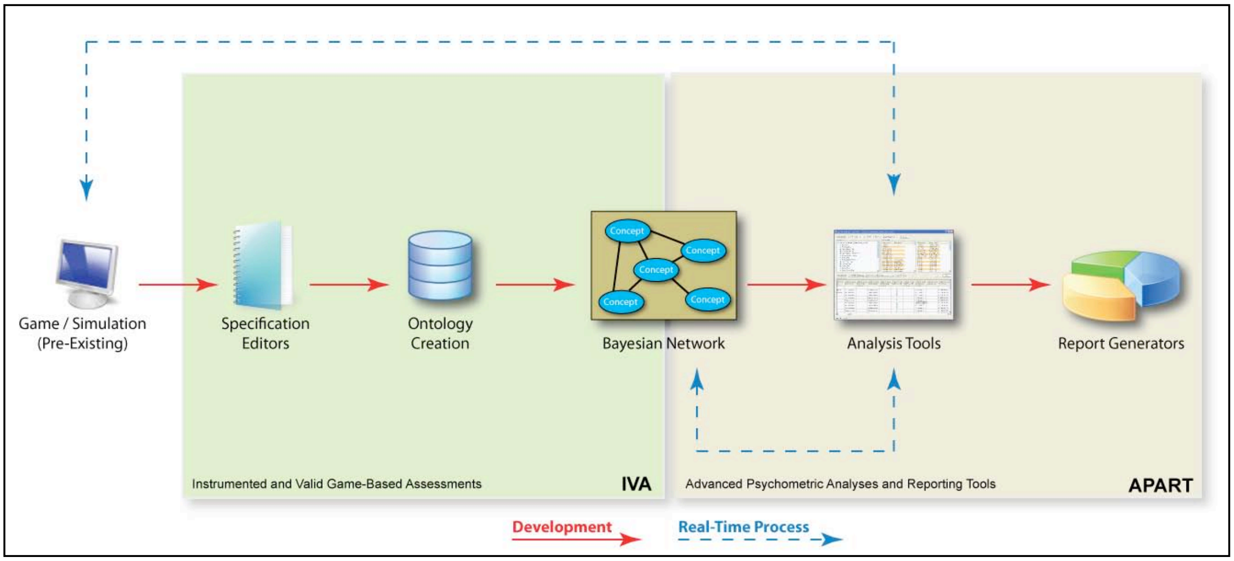


Figure 1. Conceptual Framework for Assessing Performance in Games & Simulations

FRAMEWORK OVERVIEW

Figure 1 depicts both the developmental and real-time process steps that comprise the conceptual framework that was used in our assessment process. The process flow begins with a pre-existing game or simulation that endeavors to instruct and evaluate one or more player competencies. Based on the game's design and intended use, various specification editors are used by the assessment team (in conjunction with subject-matter experts) to determine the domain (or sub-domain) that the game represents, along with the relevant tasks available in the game that are germane to the assessment.

Ontology Creation

Bounded by these specifications, an ontology is then constructed to capture the interrelationships that exist among the key concepts. Our ontology creation process draws upon pre-existing research in the field (Baker, 1998; 2007; Chung et al., 2006; Vendlinski, Baker & Niemi, 2008) and is comprised of five primary steps, as follows:

1. **Define the domain**. This is done in conjunction with subject matter experts to not only identify the broad domain being assessed (i.e. damage control onboard Navy ships), but also to tease out the relevant sub-domain that bound the players' interactions in the game (i.e. reporting and combating fire-fighting casualties onboard Navy ships).
2. **Define the ontology elements**. This step involves defining and categorizing the various elements of the ontology into one of three levels:
    a. *Top Level* – consisting of standards, big ideas, broad cognitive concepts (i.e. situation awareness, decision making, etc.).
    b. *Middle Level* – consisting of mostly unobservable (latent) variables and concepts (i.e. enemy intentions, etc.)
    c. *Bottom Level* – consisting of directly observable variables, actions, and events (i.e. using a CO_2 extinguisher, closing a valve, etc.)
3. **Create element equivalence classes.** At this step, for each element (a.k.a. object) in the ontology, we define any properties, operations, or operation rules that are relevant to that object. For example, a fire (object) can be classified as a type alpha, bravo, charlie, or delta (properties) based on the type of fuel it burns. In addition, a fire can spread, be attacked, or be extinguished (operations) based on the type of extinguishing agent used (operation rule).
4. **Define relationships within categories of objects.** Here, for each broad object represented (i.e. *Fire*), we define the relationships that exist between subordinate constituent objects (i.e. *Fire Type*, *Extinguishing Methods*, etc.). In the ontology, these relationships are expressed using phrases such as "type-of," "part-of," etc.
5. **Define relationships between categories of objects.** Finally, at this step we define relationships that exist between each of the broad objects represented (i.e. between *Fire* and *Reporting a Fire Casualty*). In the ontology, these relationships are expressed using phrases such as "property-of," "operates-on," etc.

Bayesian Network Development

The next step in our process builds upon the ontology with the aim of providing an infrastructure necessary for assessment that can effectively interpret evidence from game-play that connects to the knowledge, skills, and abilities being assessed (Shute, Ventura, Bauer, & Zapata-Rivera, 2009). It involves the construction of a Bayesian network, which is a graphical model for representing (causal) probabilistic relationships between variables. Bayesian networks have many advantages. Due to their graphical nature, they can be used to gain an understanding about a domain. They can also model such things as prior knowledge; incomplete or missing data; clean or noisy observed data; and latent, uncertain, or unobserved variables. Bayesian networks can learn both parameters and network structure from observed data, infer or predict unobserved outcomes, and they can be expanded to Dynamic Bayesian Networks (DBNs) to model time sequences of events.

The Bayesian network that gets created at this step in the process is an "operationalized" representation of the ontology. While its structure represents the same underlying relationships depicted in the ontology, its purpose is to represent the probabilities that reflect the strength of the relationship between one construct to the next. In this way, bottom level elements of the ontology (which are comprised of easily observable events) can be related through the network to top-level elements in the ontology. By constructing a Bayesian network in this fashion, we acquire the ability to assess higher order player abilities (such as situation awareness or decision making) directly by the capturing of lower level, observables arising out of game-play.

This process is highly iterative, and relies on working closely with subject matter experts to develop conditional probability tables (CPTs) that appropriately and accurately reflect the meaning and importance of player actions in the game. The goal is to represent in the Bayesian network the rules that govern an expert human rater's thought process if they were to assess a player's performance in this domain.

Analysis Tools and Reporting

The final two steps of the process involve the analyzing and reporting of data that emerges from the Bayesian network. As stated above, each observable and meaningful player action from the game can be fed into the Bayesian network to determine the probability that that action relates to one or more key constructs being assessed. In some cases, this might be a single event; in others it might be a collection of actions, the aggregate of which relates to a broader concept.

Either way, this data is then fed into an analysis tool that parses it into meaningful chunks of information, which then either get distributed back to the game, or out to a reporting engine (or both).

The idea is that this analysis tool serves as a real-time interface with the game for purposes of providing formative assessment based on player actions. For example, if a player attempts to put out a fire using an inappropriate extinguishing agent, the analysis tool can not only feedback to the game (and record) that this was an incorrect action, but can also trigger subsequent events in the game that would be contextually appropriate to remediate the player on this skill (such as providing additional fires to practice on, having a non-player character provide verbal instruction, etc.).

In addition, the analysis tool can be used to perform summative assessments of performance in which the game-play data is processed post-game to see how well particular knowledge, skills, or abilities were demonstrated. This information can then be fed directly into a reporting tool that can visually summarize the player's performance.

THE VALIDATION STUDY

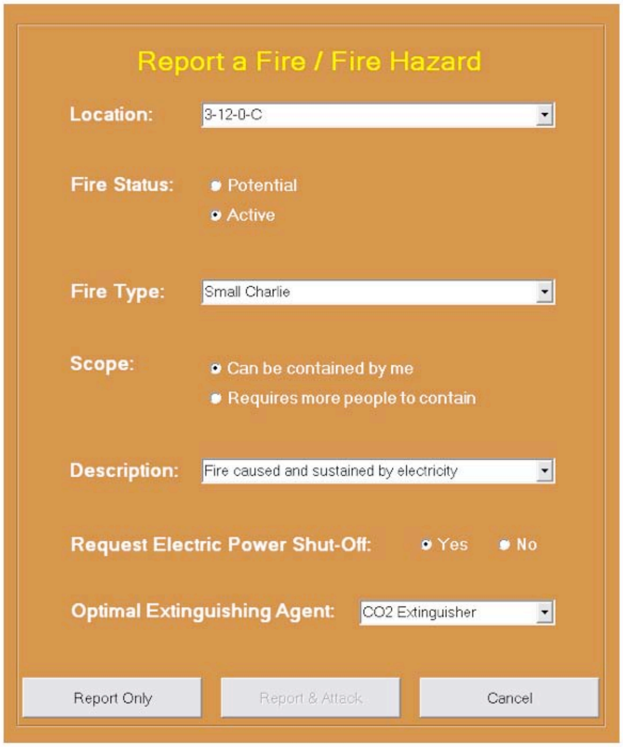
In order to validate our process, we conducted a pilot study using a preexisting damage control simulation built by the UCLA's National Center for Research on Evaluation Standards and Student Testing (CRESST). The simulation is designed to assess a player's knowledge of fire-fighting skills onboard a naval ship. The study was conducted at the Center for Naval Engineering (CNE) in Norfolk, VA., and was intended to see if a Bayesian network developed with subject matter expertise from a fire-fighting ontology could assess performance in a way that characteristically matched that of expert human raters.

Participants

Forty-five participants played through the simulation individually (35 Male, 10 female). The range of fire-fighting knowledge represented in the group was diverse, ranging from expert damage control instructors, to novice Naval Academy midshipmen with no prior fire-fighting experience or knowledge. Participants were randomly selected from various damage control classes being held at CNE.

Computer-Based Environment

The instrument used was a 3-D, first-person perspective simulation built using the Truevision3D game engine in concert with VB.Net. It was a PC-based environment that depicted the interior of a naval ship, inside of which 10 separate fires casualties existed. The player's task was to locate all 10 incidents, appropriately report them to damage control central, and if possible, attack and contain the fire using the available resources on the ship. For each incident, a reporting interface was used for the player to communicate their assessment/perception of the situation. A screenshot of the report interface is shown in Figure 2.



Report a Fire / Fire Hazard

Location: 3-12-0-C

Fire Status: ○ Potential ● Active

Fire Type: Small Charlie

Scope: ● Can be contained by me ○ Requires more people to contain

Description: Fire caused and sustained by electricity

Request Electric Power Shut-Off: ● Yes ○ No

Optimal Extinguishing Agent: CO2 Extinguisher

Figure 2. Screenshot of Report Interface

After submitting the report, the player had the option to either take no action, or use a variety of extinguishing agents to combat the fire. All player actions relating to reporting and fire containment were captured in a Microsoft (MS) Excel file, and subsequently sent to the Bayesian network for analysis.

PROCEDURE

The study was conducted in a classroom at CNE that could accommodate up to 10 students at a time. Each student was provided with a PC laptop computer on which the simulation was played. Upon entry into the classroom, the participants as a group were told of the purpose of the study, and that their participation was voluntary. They were each given ID numbers so that the data collected from their performance would remain anonymous. Each person played the same version of the simulator, and was allowed to complete it at their own pace (time was not a factor in the analysis). Assistance was only provided to address any technical difficulties that arose—all other matters were left for the student to work on unassisted.

At the conclusion of the simulation, all participants were released, and their data was automatically collected by the computer system and exported into an automatically generated

MS Excel file. Four expert fire-fighting instructors from the Damage Control School at CNE reviewed the individual MS Excel files and scored each element of each players report for all 10 incidents based on the following rubric:

- **Optimal –** best answer possible
- **Adequate –** a good answer, but an obvious better one exists
- **Poor –** correctly addresses the situation, but many better choices exist
- **Neutral –** response is unrelated to the situation
- **Bad –** response is a bad choice, and has the potential for doing more harm than good.

RESULTS

Using the Bayesian network created to analyze the responses made by the players for the report dialogs for each scenario, a scoring tool was developed to compare the player's perceptions with the known aspects of each of the ten scenarios. This was done to elicit evidence of the student's situation awareness. The conditional probability tables were populated based on expert knowledge. We analyzed a subset of the data that the experts scored using a rubric.

The observable simulation data from eight experts and seven students were recorded with a total of 40 player ratings. One rating was removed from the analysis because there was too much missing data. One expert (#1) graded all of the seven students and overall provided 14 player ratings. Another expert (#6) graded all the other experts plus one of the students, and overall provided 11 player ratings. Experts also evaluated other experts, but those numbers were small, one or five player ratings.

Demographic Data

The average age of the experts was 34 (*SD* = 6.4) and ranged from 25 to 44. All were male. Five of them were damage controlman and one was a machinist mate. The number of years in the Navy ranged from 5 to 19 years, with an average of 12.5 years (*SD* = 5.6). Seven of the experts had over 500 hours of instruction in fire fighting, flooding and casualty. Seven of the eight experts listed that they were Damage Control Leader, Fire Team Leader, Team Leader, On Scene Leader. Six of the eight listed Fire Marshal, and all eight listed that they had been an Investigator. None had been the DCA (Damage Control Assistant). On a scale of 1 (*no interest*) to 5 (*high interest*), the experts liked Action-type games the most (*M* = 3.4, *SD* = .87), then Arcade-type (*M* = 2.6, *SD* = .69) and last Real Time Strategy type games (*M* = 2.21, *SD* = .91).

The average age of the students (all from group B) was 19.4 (*SD* = 0.5) and ranged from 19 to 20. Five were male and two were female. There were all midshipman (3/c). The number of years in the Navy ranged from .83 to 2 years, with an average of 1.2 years (*SD* = 0.56). All seven students had less than 19 hours of instruction in fire fighting, flooding and casualty. One student had been an investigator on a ship, and another a fire-team leader. On a scale of 1 (*no interest*) to 5 (*high interest*), the students, like the experts, liked Action-type games the most (*M* = 3.26, *SD* = .76), then Arcade-type (*M* = 2.6, *SD* = .73) and last Real Time Strategy type games (*M* = 2.19, *SD* =1.1).

The report dialog variables (see Figure 2) for each scenario included seven elements:

- **location** (bulls eye),
- **status** (active or potential),
- **fire type** (e.g., small class A),
- **scope** (can be contained by me or requires help from others),
- **description** (e.g., fire caused and sustained by electricity),
- **optimal agent** (e.g., CO2 extinguisher), and
- **request** p**ower off** (whether a request to shut off the power was needed or not).

Response Data

Responses were saved to MS Excel files and read into the scoring tool. The tool first sets the evidence for the scenario to true (see node under the letter A on the right side of the Bayesian network diagram in Figure 3). The Bayesian Network then updates to show what is known for the scenario (see nodes under the letter B in Figure 3). The program then sets the evidence nodes (under letter C in Figure 3) to the options that the player chose and then the network is updated. The updated probabilities for the hypotheses nodes (see nodes under the letter D in the center of Figure 3) are then mapped to performance levels using a lookup table.

Finally, in MS Excel, we compared the Bayesian Network rating to the expert ratings. The exact agreement percentage can be found in Table 1. The table also includes the exact agreement percentage among the Bayesian Network determination of the overall reporting performance with the expert's overall rating. The Bayesian Network determination is a formula that has a weighting based on the following formula (1):

(64*Location_RepVsReal+32*FireType_RepVsReal+16*OptimalAgent_RepVsReal+8*SecurePower_ RepVsReal+4*Scope_RepVsReal+4*Status_RepVsReal+4*Description_RepVsReal)/132 (1)

The numbers in this formula that precede each of the report elements represent the relative weighting of importance of the item compared to the other items of the report. In the formula, the "RepVsReal" are the updated probabilities from the Bayesian Network that compare the reported (Rep) versus reality (Real). The relative importance of the report elements was elicited through expert consensus.

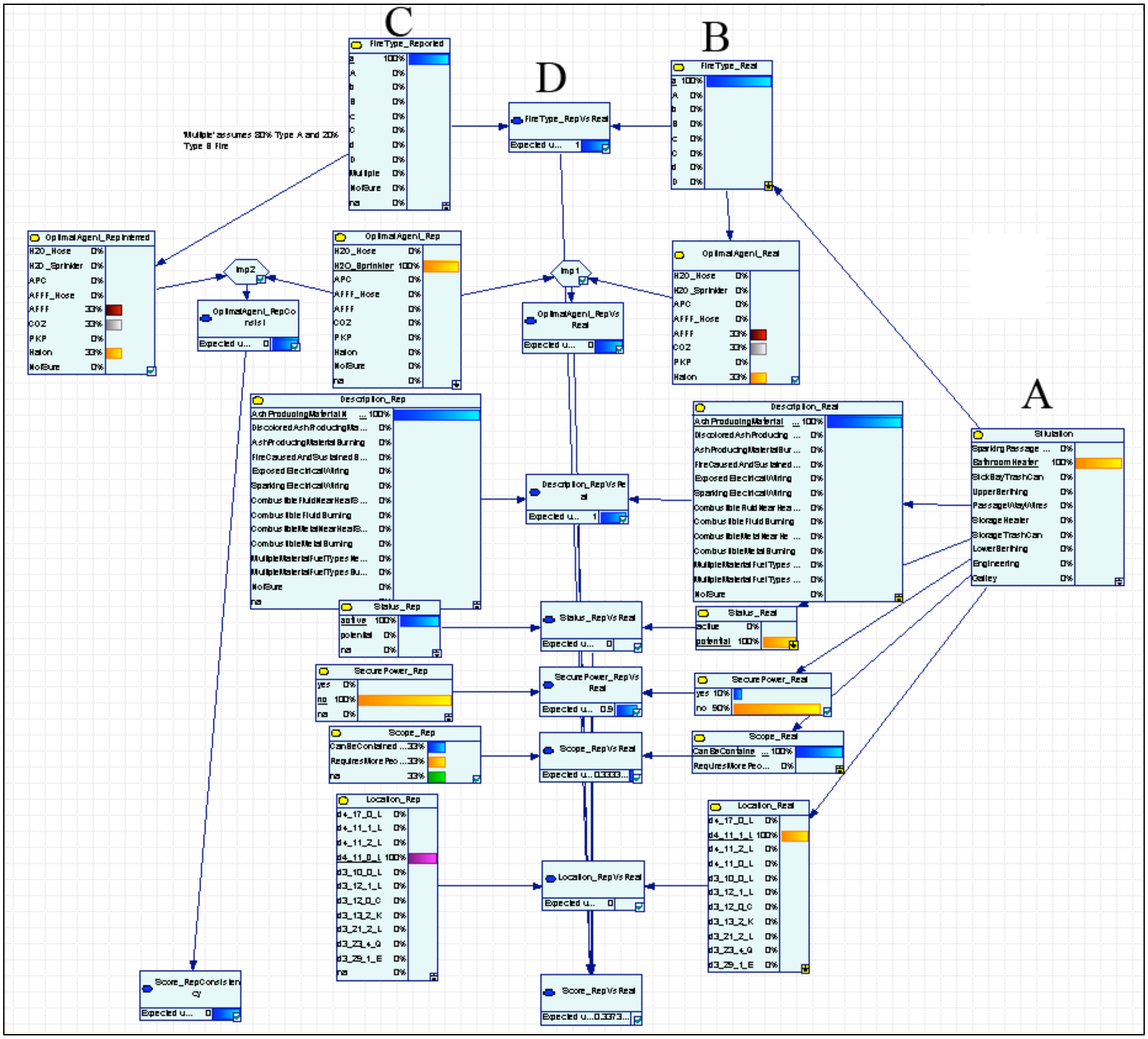


Figure 3. Bayesian Network for Report Dialogs

Table 1.

Percentage of Agreement between Bayesian Network Scoring and Expert Scoring

Scenario	Location	Status	Firetype	Scope	Description	OptAgent	PowerOff	Report Overall
Bathroom Heater	100	100	92.3	97.4	79.5	48.7	7.7	67.6
Engineering	100	100	97.4	56.4	71.8	74.4	94.9	78.4
Galley	100	100	82.1	100	50	55.6	21.1	18.9
Lower Berthing	100	100	21.1	87.2	17.9	63.2	71.8	43.2
Passage Way Wires	100	100	66.7	38.5	87.2	69.2	82.1	44.1
Sick Bay Trash Can	100	100	94.9	100	64.1	71.8	79.5	80.6
Sparking Passage Way Panel	100	66.7	97.4	7.7	48.7	100	100	78.9
Storage Heater	100	100	76.9	100	82.1	51.3	74.4	31.6
Storage TrashCan	100	100	97.4	100	82.1	64.1	82.1	73
UpperBerthing	100	100	97.4	87.2	89.7	31.6	5.1	60
Average	100	96.67	82.36	77.44	67.31	62.99	61.87	57.63

Note. OptAgent = optimal extinguishing agent.

The match was highest for the location, then status, firetype, scope, description, optimal extingusihing agent, power off, and overall reporting. These results suggest that the model needs refinement and/or that some scenarios may have been ambiguous.

SUMMARY AND DISCUSSION

Approximately 8 hours were spent working with 10 different subject matter experts from the CNE fire-fighting school to facilitate the creation of the Bayesian network. This activity involved one-on-one interactions as well as group discussions, the result of which yielded consensus among all experts on how to score each of the 10 scenarios depicted in the game. Despite this, however, as Table 1 shows, significant disagreement exists between the Bayesian network and the expert scoring.

There are several reasons why this may have occurred. The first has to do with differences in how humans access and retrieve knowledge compared to computers. When people encounter a situation to evaluate, they attempt to comprehend it in terms of existing scripts (or schemas) they already posses about similar past experiences (Schank, 1999; Ratcliff & McKoon, 1988; Wattenmaker, 1992). These scripts are recalled and understood not based solely on the factual content that comprises the situation, but on the storied context that integrates these facts with particular experiences (Ferguson, Bareiss, Birnbaum, & Osgood, 1992). As such, experts often possess a lot of implicit knowledge that colors their understanding of a situation, and which can be difficult to articulate in words.

When working with the subject matter experts to encapsulate their knowledge into the relationships and conditional probability tables of Bayesian network, it is likely that nuances of how they would evaluate a player's performance were excluded because this information was difficult to ascertain in the absence of a specific case to analyze. As a result, the Bayesian network was not robust enough to appropriately score a player's performance under certain circumstances.

Another possible reason for the discrepancy in scoring has to do with the consistency with which the experts adhered to the agreed-upon scoring rubric. Despite reaching consensus on how each scenario should be scored, the scoring took place over several days, and therefore the experts might not have remembered all the conventions agreed upon when they actually performed the player evaluations. Furthermore, the experts were often interrupted with other job-related tasks they needed to perform, resulting in distracting lapses in time when scoring even a single player.

All of these reasons underscore the notion that although the conceptual process of creating and training a Bayesian network to assess performance is fairly straightforward, successfully implementing it where it reliably replicates human scoring can be very difficult. Indeed, this small-scale validation study exemplifies the fact that this process is highly iterative, and that even for relatively simple scenarios, the wide variety of player responses poses a daunting challenge for devising a robust Bayesian network.

Next Steps

The next steps to this project include further refining the Bayesian network with expert input, and then to score player actions undertaken to combat the fires they encountered in the game. This poses an order of magnitude increase in complexity over the current phase of just evaluating reports of fire casualties, as both situation awareness (arising largely from what the player reports) and decision making (arising from the specific fire containment actions taken) will collectively be considered in the final scoring.

Once this work is completed, the network will be capable of moving beyond its current state (in which formative and summative assessments can be provided for the reporting of fire casualties only) to being able to fully assess situation awareness and decision making as it pertains to the entire fire-fighting process. We anticipate achieving this capability by December 2009.

REFERENCES

Baker, E. L. (1998). *Model-based performance assessment* (CSE Tech. Rep. No. 465). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L. (2007). *Moving to the next generation system design: Integrating cognition, assessment, and learning* (CSE Tech. Rep. No. 706). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, K. W. K., Baker, E. L., Delacruz, G. C., Elmore, J. J., Bewley, W. L., & Seely, B. (2006). *An architecture for a problem-solving assessment authoring and delivery system*. (Deliverable to the Office of Naval Research). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

Ferguson, W., Bareiss, R., Birnbaum, L., & Osgood, R. (1992). ASK systems: An approach to the realization of story-based teachers. *The Journal of the Learning Sciences, 2*(1), 95–134.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective, 1*(1), 3–62.

Pascual, R., & Henderson, S. (1997). Evidence of naturalistic decision making in military command and control. In C. E. Zsambok, & G. Klein (Eds.), *Naturalistic decision making*, (pp. 217–226). Mahwah, NJ: Lawrence Erlbaum Associates.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*(3), 385–408.

Schank, R. C. (1999). *Dynamic memory revisited*. New York, NY: Cambridge University Press.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects,* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

Vendlinski, T. P., Baker, E. L., & Niemi, D. (2008). Templates and objects in authoring problem-solving assessments (CRESST Report 735). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wattenmaker, W. D. (1992). Relational properties and memory-based category construction. *Journal of Experimental Psychology, 18*(5), 1125–1138.