Automated Assessment in Games & Simulations

2015 CRESST Conference

Alan Koenig, Ph.D.

August 19, 2015
Automated Assessment

What is Automated Assessment?
A software-based solution to assessing a student’s knowledge, skills, and/or abilities (KSA’s)

Why use automated assessment?
Games & simulations often require students to engage in cognitively complex tasks, which can involve:
- multiple, non-trivial steps to complete
- interdependent tasks
- multiple pathways to success

Which means… assessment can be difficult:
- Detecting meaningful actions
- Making sense of meaningful actions (what, when, how, under what conditions)
- Inferring latent KSA’s from observed actions/events

Automating this process can help clarify and validate human judgment.
The Challenge: Getting what you want out of a game or simulation

Data of Observable Actions/Events:
- What happened?
- When?
- In what context?
- Who acted?
- With what resources?
- Etc.

But often what we want to know about a student is NOT directly observable. Instead, it has to be inferred.
You did well in the sim... Now what?

You accrued 136 points, and got a rating of “Proficient”

But what I really want to know is: will you be safe piloting a DDG in a crowded, chaotic harbor?
Automated Assessment Methodology

**GOAL:**
What do we want to know about this person?

**CONTEXT:**
In what domain do we want to assess? Under what conditions?

**RELEVANCY:**
What are all the concepts, facts, procedures, and skills this person should know to show proficiency?

**EVIDENCE:**
What observable, *meaningful actions / events* should we pay attention to?

---

I want to assess because…

This is what’s relevant…

This is how I infer based on what I see…

Instructor / SME

National Center for Research on Evaluation, Standards, & Student Testing
Automated Assessment Methodology

Example: Shiphandling Mooring Task:

What’s the assessment goal?
Determine a student’s shiphandling proficiency with mooring a DDG (Navy Destroyer) to a pier.

What latent knowledge, skills, and abilities (KSA’s) are of interest?
- Proficiency with maneuvering a ship (use of rudder / propulsion / tugs)
- Ability to maintain safe practices

Under what conditions do we want to assess?
- Port of Bahrain, mooring to occur between two docked ships
- Offsetting, light wind & current
- Light harbor traffic
- Daytime
Automated Assessment Methodology

- Identify Assessment Specifications
- Develop Ontological Mapping
- Develop Bayesian Network

Ontologies are visual representations of domain knowledge:

- Nodes are elements (procedures, tasks, skills, concepts, etc.)
- Lines are the relationships among elements

Ontologies aid instruction and assessment:

- Support scenario development
- Support evaluation of performance

Ontologies aid instruction and assessment:

- Support scenario development
- Support evaluation of performance

National Center for Research on Evaluation, Standards, & Student Testing
Ontology Construction
Mooring Ontology (zoomed, partial view)
Automated Assessment Methodology

Steps:
1. Identify meaningful actions / events to detect & observe
2. Construct the Bayesian network
3. Develop scoring rubrics for meaningful actions / events
Identify Meaningful Actions

Based on the assessment goals and ontology, a set of observable, meaningful actions are identified. It is from these actions that inferences to latent skills are made.

**Shiphandling Mooring: Meaningful Actions to Consider**

- Clearance to Buoys
- Clearance to Other Ships
- Ship’s Heading Steadiness
- Ship’s Heading from Pier Heading
- Speed Over Ground (fore-aft)
- Speed Over Ground (lateral)
- Approach Track (*within Green Zone*)
- Docked “Bridge-Here” Alignment

What experts pay attention to
**Construct the Bayesian Network**

- Constructed using information from the ontology
- Links represent dependencies between ontology parts. They indicate strength of relationships (stored as conditional probabilities)
- Probability of mastery of the latent variables is inferred from observable actions

**Everything’s expressed as a probability:**

- **P(PS):** Probability of mastery of skill Latent Variable 3
- **P(E | PS):** Conditional probability. Probability of mastery of skill Observable Action 4, given information about mastery of skill Latent Variable 3
- **P(E):** Probability of mastery of concept Observable Action 4

\[
P(PS | E) = \frac{P(PS) \times P(E | PS)}{P(E)}
\]
Shiphandling Mooring Bayesian Network

- Maneuver
- Safety
- Rudder/Prop/Tugs

Observable (Meaningful) Actions:
- Heading Steadiness
- Approach Track
- Docked Alignment
- Clearance
- S.O.G (lateral)
- S.O.G (fore-aft)
- SH vs. PH
- Approach Track

Latent Variables
Develop Rubrics for Observable Actions

- Each observable action (node) in the Bayesian network is "evaluated" using a rubric.

- Depending on the node, the evaluation can be triggered by:
  - an action / event (i.e. a collision event)
  - a time interval (i.e. evaluate ship’s heading every 20 seconds)

- The result of each evaluation is a score, which is fed into the Bayesian network.

**Example: Ship Heading from Pier Heading Rubric**

![Ship Heading from Pier Heading Rubric](image)

National Center for Research on Evaluation, Standards, & Student Testing
Bayesian Network (Inferences) Update in Real-Time

Updates nodes here

Evidence supplied here...
The Automated Assessment Engine in Action

Simulator Components:

1. Conning Officer Virtual Environment (COVE)
   A high-fidelity shiphandling simulator used by the Surface Warfare Officer’s School (SWOS) in Newport, RI.

2. Intelligent Tutoring System (COVE-ITS)
   Provides spoken coaching to student based on observed actions.

3. Automated Assessment Engine
   Evaluates observed (meaningful) actions, and infers latent skills of student
COVE Assessment: System Architecture

- **Raw Telemetry (1 Hz)** (student actions/events)
- **Filtered Telemetry (1 Hz)** (meaningful actions)
- **Spoken Coaching**
- **Inferred Skills**

**COVE** - ITS

**Automated Assessment Engine**

**Student**
Validation: Does AAE Match Expert Conclusions?

Subjective Scoring

Match??

Automated Assessment Engine

Bayesian Network Analysis

What really matters to experts?

National Center for Research on Evaluation, Standards, & Student Testing
## Validation: Does AAE Match Expert Conclusions?

Reliabilities of Scoring between Master Mariners and Bayesian Network (BN)

<table>
<thead>
<tr>
<th>Sample Size: n = 9</th>
<th>Cronbach alpha</th>
<th>Krippendorff alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 master mariners</td>
<td>2 mm +BN</td>
</tr>
<tr>
<td>Maneuver</td>
<td>0.952</td>
<td>0.897</td>
</tr>
<tr>
<td>Safety Margins</td>
<td>0.978</td>
<td>0.869</td>
</tr>
<tr>
<td>Use of Rudder, Propulsion, and Tugs</td>
<td>0.943</td>
<td>0.883</td>
</tr>
<tr>
<td>Clearance</td>
<td>0.326</td>
<td>0.463</td>
</tr>
<tr>
<td>Ship Heading from Pier Heading</td>
<td>0.906</td>
<td>0.631</td>
</tr>
<tr>
<td>Approach Track</td>
<td>0.662</td>
<td>0.703</td>
</tr>
<tr>
<td>Speed Over Ground (fore-aft)</td>
<td>0.604</td>
<td>0.539</td>
</tr>
<tr>
<td>Speed Over Ground (lateral)</td>
<td>0.677</td>
<td>0.609</td>
</tr>
<tr>
<td>Heading Steadiness</td>
<td>0.632</td>
<td>0.569</td>
</tr>
<tr>
<td>Bridge Here Alignment</td>
<td>0.938</td>
<td>0.931</td>
</tr>
</tbody>
</table>

- **Latent (inferred) skill mastery**
- **Implicitly scored by instructors using Navy rules/rubrics**
- **Explicitly scored by BN based on Navy rules/rubrics**
National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

VISIT US ON THE WEB
cresst.org

Copyright © 2013 The Regents of the University of California