# Latent Ability Estimation in Serious Games Through Gameplay Data Visualization

Anna Lee, Thomas Maierhofer
UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

## Introduction

This project estimates latent traits from students' behavior in a serious game, which is a game designed for a primary purpose other than pure entertainment (Loh, Sheng, and Ifenthaler 2015). The game used in this study was designed by CRESST to assess deductive reasoning skills. In the two levels (referred to as Scenario 6 and Scenario 7) used to assess reasoning, the game asks students to navigate a UFO through a 2D maze in darkness with only the help of a flashlight, see Figure 1. Their ability to successfully navigate the maze grants insight into their deductive reasoning ability. The study provides a useful guideline for assessing player performance in real-time by introducing a procedure that transforms unintelligible game logs into visual representations.

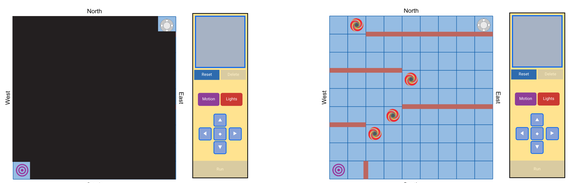


**Figure 1.** Scenario 7 as the player sees it (left) and with lights on (right). A version of this game with other scenarios is available for testing at ufo.cresst.net.

## Game Design

The goal of the game is to move the UFO to the target without hitting any obstacles: vortexes, which can be jumped over, and walls, which cannot. If a player runs into an obstacle, the attempt is restarted. Students are unable to see the walls and vortexes unless they use the flashlight to look ahead. They operate the UFO through a remote control with directional (North, East, South, and West) and action buttons (Motion, Light, and Jump). The game was administered to 19 12th grade students from which the telemetry data was collected. They also were given a five-question multiple choice test as an external reasoning measure to validate the data against. Each student was rated 0-5 based on the number of questions he or she answered correctly.

## Mapping Actions

Each player's actions in their attempts are plotted on the game grid and color-coded according to the action category. These maps visually demonstrate various player strategies based on action types utilized during the game, as well as how they learned over time, see Figure 2. For instance, a player might initially try a trial-and-error method to move through the maze, then learn to use the light and make informed moves. The maps provide insight into the player's thought processes.

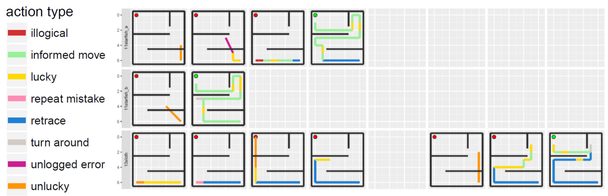


**Figure 2.** Each row is a new player and each column is a different attempt. The target is represented by a circle colored green for success on the attempt, red for failure. The black lines represent wall obstacles. In the empty map, the player hit the reset button immediately upon starting the attempt, resulting in a new attempt.

## Categorizing Actions

Logging all player actions results in a multivariate time series of the player actions' mode (Motion, Lights, Jump), direction (North, East, South, West), and the initial and final position of the UFO (x and y coordinates). This automatically collected information is human-readable, but high-level strategies cannot be identified from the raw data.

In a first step, all actions are categorized into one of nine comprehensive and disjunctive action categories by walking through the binary decision tree depicted in Figure 3. This tree outlines a function for automatically labeling every player action. A lights action can be "informative use" or "redundant" depending on whether or not new information was gained. A motion action can be "lucky," "unlucky," "informed," "illogical," "repeat mistake," "turn around," or "retrace." The distinctions within motion actions are based on whether the outcome of the move is known to the player, it has been previously done or seen with the light, it has been done in this attempt or a previous attempt, or it led to a crash. These criteria are automatically computed from the log by creating a model for every player's knowledge and previous moves that is updated after every move, seen in Figure 4.

**Figure 3 (right).** Binary decision tree which summarizes the action categories. Every action is categorized by walking through the binary decision tree from the top down until a leaf node is reached. This tree is a visualization of the function that automatically labels each move players made in-game.

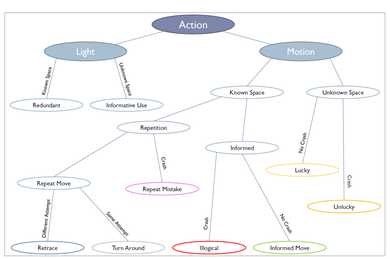


**Figure 4 (bottom).** Snapshot of the data log used to generate a model for every player's knowledge and moves. By evaluating an action against the knowledge a player already has, latent cognitive abilities can be inferred.

| attempt | Success | action | direction | start.xloc | start.yloc | end.xloc | end.yloc | type |
|---|---|---|---|---|---|---|---|---|
| 8 | TRUE | MOTION | SOUTH | 0 | 1 | 0 | 2 | retrace |
| 8 | TRUE | MOTION | EAST | 0 | 2 | 1 | 2 | retrace |
| 8 | TRUE | MOTION | EAST | 1 | 2 | 2 | 2 | retrace |
| 8 | TRUE | LIGHTS | EAST | 2 | 2 | 2 | 2 | informative use |

## Bayesian Network

The nine action categories are scored on a scale for reasoning ability and risk affinity by two expert raters following Shute, Ventura, Bauer & Zapata-Rivera (2009). High agreement (Cronbach's alpha > 0.95 for both deductive reasoning ability and risk affinity) allows for averaging their scores. These scores are used to train a Bayesian network predicting action category based on latent reasoning ability and risk affinity. This network allows the probability of the actions given a combination of latent traits (Figure 5) and the estimation of the latent traits given an action (Figure 6).

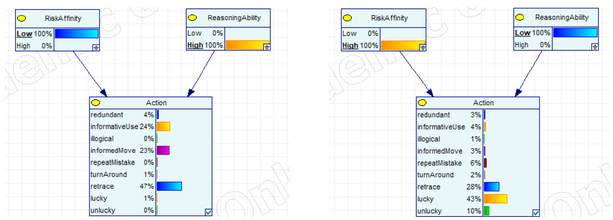


**Figure 5.** Bayesian network predicting action category. Students with low risk affinity and high reasoning ability are expected to have more informative light use and take more informed moves (left). Students with high risk affinity and low reasoning ability are expected to crash and make uninformed moves more (right).

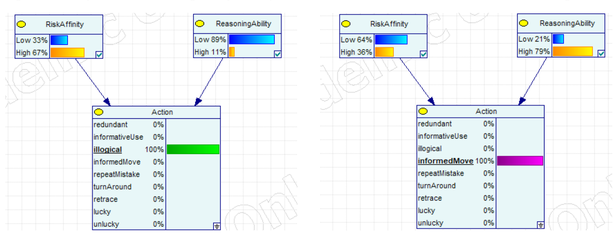


**Figure 6.** Bayesian network predicting latent traits. An illogical move (crashing even though the player should know it would happen) indicates a high risk affinity and low reasoning ability (left). An informed move (moving into a space that is known to be safe) indicates a low risk affinity and high reasoning ability (right).

## Estimation of Latent Traits

Applying this Bayesian network onto every student's sequence of actions results in a player profile of estimated risk affinity and reasoning ability over time (Figure 7). It is not a dynamic Bayesian network, so it explicitly neglects the temporal order of the observations. The non-dynamic Bayesian network's theoretical independence assumption of latent traits at action t and action t+1 is obviously violated. A credible estimation of the latent traits at the time of action t is achieved by non-linear smoothing where a loess estimator is used to produce a smooth estimate (Figure 7). Neglecting the temporal order in the Bayesian network enables a more flexible estimation of the temporal development of latent traits.

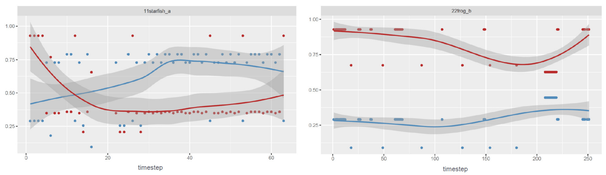


**Figure 7.** Estimated risk affinity (red) and reasoning ability (blue) for every action performed to complete one scenario and smooth line with 95% confidence interval. The action category "retrace" (repeating a move after a crash to get back) was omitted due to its high occurrence and low informativeness. The left figure shows the profile of a player who is quickly adapting a successful strategy (initial high risk affinity and low reasoning invert after a while), while the right figure shows a player who is mostly guessing (way more actions overall, high risk affinity and low reasoning ability throughout).

## Linear Regression

As an external validation of the estimated latent traits, the average estimated latent reasoning ability was compared to students' performance on a paper-based reasoning test taken before playing the game. In both scenarios, the correlation coefficients suggest substantial agreement of the reasoning ability estimated in-game and on paper. As expected from the Bayesian network results, risk affinity had an inverse correlation with reasoning ability. It was not assessed externally as a separate measure.

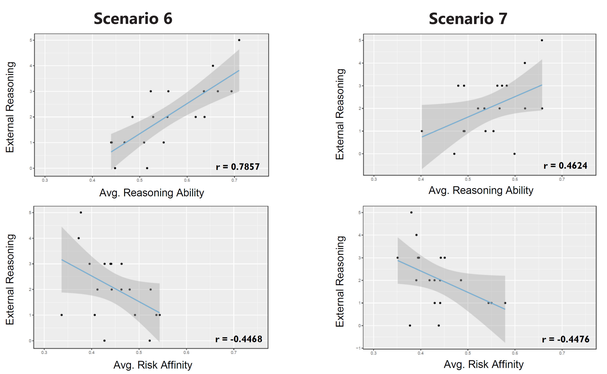


**Figure 8.** The average reasoning ability and risk affinity scores from the Bayesian network were compared against the external reasoning measure.

## Conclusion

The estimated reasoning ability shows substantial agreement with a paper-based reasoning test, supporting the validity of the proposed assessment. The project provides a guideline for visualizing player performance in real-time, with potential applications in player assessment and providing adaptive in-game feedback. Further validation of this work is in progress.

## Contact

**Anna Lee** alee14@ucla.edu
**Thomas Maierhofer** maierhofer@cresst.org

## References

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). Serious Games Analytics. *Edited by Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler. Cham: Springer International Publishing. doi, 10*: 978-3.

Shute, Valerie J, Matthew Ventura, Malcolm Bauer, and Diego Zapata-Rivera. 2009. "Melding the Power of Serious Games and Embedded Assessment to Monitor and Foster Learning." *Serious Games: Mechanisms and Effects, 2*: 295–321.