A Data-Driven Approach for Inferring Student Proficiency from Game Activity Logs

Mohammad H. Falakmasir[⋆,#] **José P. González-Brenes**[⋆] **Geoffrey J. Gordon**[§] **Kristen E. DiCerbo**[⋆]

[⋆]School Research
Pearson
{jose.gonzalez-brenes,
kristen.dicerbo}@pearson.com

[#]Intelligent Systems Program
University of Pittsburgh
falakmasir@pitt.edu

[§]Machine Learning
Carnegie Mellon University
ggordon@cs.cmu.edu

ABSTRACT
Student assessments are important in education because they allow collecting evidence about student progress. Unfortunately, they can be very tedious to the stakeholders. We present a novel machine learning algorithm, Student Proficiency Inferrer from Game data (SPRING), that allows modeling game playing behavior in educational games. Unlike prior work, SPRING is a fully data-driven method that does not require costly domain knowledge engineering. We validate our method using data collected from students playing 12 educational mini-games. Our results suggest that SPRING is accurate to predict Math assessments (R^2 =0.55 , Spearman ρ=0.82).

ACM Classification Keywords
K.3.1 Computer Uses in Education.

Author Keywords
Educational Games, Unsupervised Learning, Invisible Assessment

INTRODUCTION
Educational assessments are important because they collect evidence about about whether the teaching goals are being met. Unfortunately, the process of administering assessments is usually disconnected to the learning environment, and it is often disruptive to the classroom. In many developed countries, students now find themselves spending increasing amounts of time preparing and taking tests instead of learning [6]. For example, a survey of the current state of testing in America revealed that students are taking an average of 113 standardized tests between pre-K and highschool [7]. For these reasons, it is not surprising that the recent political climate and the general population have been weighing in on the question of whether students are being tested too much [7].

According to Evidence Centered Design (ECD) [12], the goal of assessment is to characterize the strength of evidence regarding claims one wants to make about individuals or groups.

Therefore, the assessment process involves identifying, organizing, or creating activities for students so we may observe that evidence. An interesting alternative to traditional (summative) assessment is invisible assessment [13], where the evidence is gathered from leaners unobtrusively from the digital interactions of their ongoing activities. This data is used to understand claims regarding what students know and what they can do [14].

A promising opportunity for invisible assessment is using log data collected from educational games. Unfortunately, engineering a system that parses logs is costly and time-consuming. For example, prior work [13, 14] has relied extensively on subject matter expertise to define features for a regression model confirm if it's regression give examples of the things they did manually .

Invisible assessment from game data may become more accurate and cheaper to implement if the domain knowledge engineering could be automated by a data-driven process. Game data is often logged with what we call a *slot and filler* structure. In Table 1, we show a simplified example of a real educational game log that uses slot and filler structure. The slots are discrete sets of events that are initiated by the leaner. Each slot may accept zero to multiple fillers. Each filler represents a value of a property of the slot event. For example, a "move" event that represents that the learner moved an object in the screen may have an *x* and *y* fillers, to represent the target position in two-dimensional space. Unfortunately, generic machine learning algorithms input *feature vectors* or *sequences* and not slot and filler structures. Describe what are feature-vectors methods and sequences .
Traditional machine learning algorithms, and off-the-shelf data mining packages, such as Weka (cite) , cannot help use slot and filler data because complete. describe why each (e.g, regression trees / HMMs) cannot be u

Table 1. An example fragment of a log from an educational game.

Id	User Id	Event Name	Event Data
1	ABC	Game Start	{ }
2	ABC	Move Object	{endPosX:363,endPosY:82}
3	ABC	Move Object	{endPosX:361,endPosY:54}
4	ABC	Insert Ruler	{EndOffsetX: -14}
...			

In this paper, we propose *Student PRofficiency INferrer from Game data* (SPRING), a novel machine learning algorithm that models game playing behavior. SPRING allows modeling raw data in slot and filler structure. We demonstrate our approach on real student game playing data. Our experiments suggest that SPRING can be used to predict student proficiency without costly domain expertise.

SPRING FRAMEWORK

Algorithm 1 The SPRING algorithm.

Write
if $i \geq maxval$ **then**
  $i \leftarrow 0$
else
  if $i + k \leq maxval$ **then**
    $i \leftarrow i + k$
  end if
end if

In this section, we describe the discretization, sequence modeling and regression steps of Algorithm 1.

Discretization

In order to gather the evidence for learning from student trajectories of interactions with the game, we needed to create fearures that are homogenious and can be used in an interpretable model. Each student trajectory consists of a sequence of discrete actions accompanied by some details that are usually continuous variables. For example, for each movement there is a two dimensional coordinates of the target x and y and for each instance of using the gluumi tool do not talk about Alice in WL here? , there is a variable that represents the size of the new object that has been created. Our main challenge in this phase was to transform the movements into a set of observable values that are usefull as evidence for learning. In order to make sure that the model is generalizable across multiple game levels, we first used DBSCAN DBScan is an example, not the only way to do this clustering algorithm [4] on the x and y coordinates of the movements to extract the crucial regions within the game interface. Then we transformed each movement into its cluster number and considered outliers as a separate cluster. use images for examples

Sequence Modeling

In the modeling phase, we hypothesized that high- and low-performing students have similar usage patterns that are representative of their sequential decision-making process. Our model should be able to successfully capture this difference across different game levels and use it to predict post-test scores. Hidden Markov Models are good candidate for the task of unsupervised analysis of sequential data. Given the sequence of student actions, we aim to infer meaningful sequence of (latent) states, which describe the process that generated the actions, along with statistical patterns that can describe and distinguish those states. need to describe more

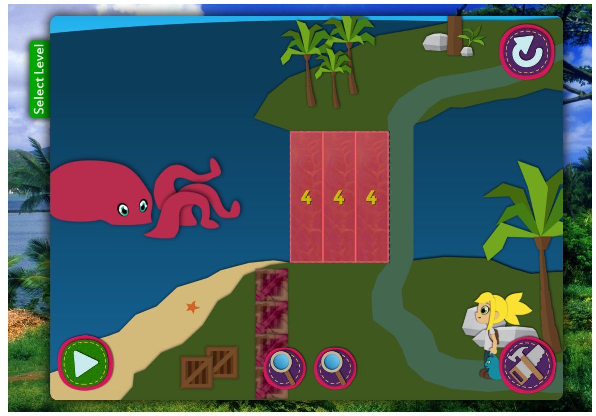

Figure 1. A screenshot of hint provided in game level 11, *You Kraken Me Up!*, in *Alice in AreaLand*. Students should combine four squares into a column and create three copies of the column to cover the designated area and prevent the octopus from attacking *Alice* while she crosses the bridge.

Regression

Write

EMPIRICAL EVALUATION

Game Environment

Alice in AreaLand is an educational game developed for research purposes. It focuses on teaching and assessing geometric measurement, specifically the understanding of area, among 6th grade students. The game targets three main stages in the development of area: 1) area unit iteration, 2) use of unit squares to measure area, and 3) use of composites to measure area. The current version has 12 game levels. A simple student scenario involves covering a 2D area with smaller unit squares placed end-to-end in non-overlapping fashion, combining the single squares into rows or columns, and then determining the number of rows or columns needed. Figure 1 shows a screenshot of one game level.

Throughout the game, *Alice* is accompanied by *Flat Cat* – an assistant character who provides feedback and scaffolding to the player in the beginning of each game level and upon request when students push a hint button (represented by two magnifiers at the bottom center of Figure 1). Earlier game levels are designed for students to learn about area unit iteration and usually require them to cover a number of predefined areas with unit squares (not necessarily in a non-overlapping fashion). By advancing through game levels, students are presented with three tools: *Gluumi* for combining unit squares by gluing them together; *Multy*, for making copies of different objects; and *Esploda* for breaking compound shapes into single units. There is no limit for completing a game level regarding time or number of actions students may execute. The students press the *Go Alice* button (bottom left corner of Figure 1) if they deem their performance to be satisfactory for *Alice* to proceed. Based on the covered area and the arrangement of the tiles, they either advance to the next level or receive a feedback and stay in the same level.

Dataset

Our dataset consists of time-stamped interactions of 129 students in 12 game levels. For 77 students, we also have post-test scores from a paper-based exam with 20 questions in the 3 skills of geometric measurement. In total, there are 88,458 events recorded in the dataset from 1,510 game sessions, meaning that student tried some of the game levels for multiple times. Based on the ECD framework, beginning levels only involve area unit iteration skill and the other skills and related features are gradually added to the later game levels. We only used the trajectories of the students who participated in the post-test. In the case of multiple attempts in a game level, we only considered the trajectory from the first attempt.

histogram of events descriptive statistics

Experimental Setup

First we divide students into two groups, one (%80) for training and development purposes and the other (%20) for test and verification. In the synthesis phase, we transform log data of different nature (eg. movements, use of the tools, requests for hints) in each game level to observable values that can be used as evidence for learning. In the modeling phase, we use the observations to train two Hidden Markov Model (HMM)s that capture the sequence of actions for high- and low-performing students. In the aggregation phase, we uses the likelihood of students' sequence of action in order to build a regression model that is predictive of the post-test results. Finally, in the verification phase, we test the regression model on the held-out set (%20) and report the results.

Results

write

RELATION TO PRIOR WORK

In order to deal with such highly unstructured data, researchers often use carefully designed network structures (such as Bayesian Networks [1]) or game-specific heuristics and benchmarks generated by experts playing the game [11, 15]. However, this approach is extremely labor intensive and might fail to capture meaningful patterns in student exploratory habits within the game. Given these limitations, data-driven analysis of student interactions provides a powerful alternative that facilitates the discussions around what does and does not work in a particular educational game.

The potential of computer games for educational purposes has been of interest since nearly the beginning of videogames. Unlike video games, which focus on creating an entertaining experience for the user, educational games require principles and strategies that engage students while maximizing their learning gain. Therefore, data-driven analysis of student behavior is crucial to better understand the learning process and improve the tools in the future.

There have been numerous attempts among the educational research community to develop analytic methods and build predictive models based on the data from educational games. *Rumble Blocks* is an educational game designed to teach basic concepts of structural stability and balance to children in grades K-3 (ages 5-8 years old). Harpstead et al. [5], studied the alignment of game to its target learning goals by examining whether student solutions follows the targeted principals. They employ clustering techniques on the individual solutions created by actual students and use principle-relevant metric (PRM) to measure how closely the representative solution embodies a specific targeted principle. The results demonstrated a misalignment between the feedback provided to students and the targeted knowledge.

Battleship Numberline is another educational game for understanding fraction using number line estimation. Students attempt to explode target ships and submarines by estimating numbers on a number line. Lomas et al. [10] performed a large-scale online experiment in order to study the effect of challenge on player motivation and learning. They presented different configurations of the game for different groups of students and used a combination of time spent and challenges attempted as a measure of engagement and the average success rate of each design configuration as a measure of challenge. The results showed a linear correlation between challenge (difficulty) and engagement, meaning the easier the game, the longer students played.

Refraction is another educational game for learning about fractions by splitting laser beams into fractional amounts to target spaceships by avoiding asteroids. Liu et al. [9], created an ensemble algorithm that combines elements of Markov models, player heuristic search, and collaborative filtering techniques with state-space clustering in order to predict player movements on last game-level based on the history of movements in previous game levels. Lee et al. [8] extended the former framework by building state-action graph and using feature selection techniques to reduce the number of features for each state. To ensure extensibility, they also tested the framework on another game *DragonBox* and reported improvement over a Markov predictor.

CONCLUSION

Limitation, we don't compare to Valerie Schultz

Limitation, we don't control for game ability like Valrerie does

Limitation, we only used data from Alice in wonderland

Method is accuaate

REFERENCES

1. David W Albrecht, Ingrid Zukerman, and An E Nicholson. 1998. Bayesian models for keyhole plan recognition in an adventure game. *User modeling and user-adapted interaction* 8, 1-2 (1998), 5–47.
2. Russell Almond, Linda Steinberg, and Robert Mislevy. 2002. Enhancing the design and delivery of assessment systems: A four-process architecture. *The Journal of Technology, Learning and Assessment* 1, 5 (2002).
3. KE DiCerbo and JT Behrens. 2012. Implications of the digital ocean on current and future assessment. *Computers and their impact on state assessment: Recent history and predictions for the future* (2012), 273–306.

4. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

5. Erik Harpstead, Christopher J MacLellan, Vincent Aleven, and Brad A Myers. 2014. Using extracted features to inform alignment-driven design ideas in an educational game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3329–3338.

6. Peter Hofman, Bryan Goodwin, and Stuart Kahl. Re-Balancing Assessment: Placing Formative and Performance Assessment at the Heart of Learning and Accountability. (????).

7. M Lazarín. 2014. Testing overload in Americas schools. *Center for American Progress. Retrieved from www.americanprogress.org/issues/education/report/2014/10/16/99073/testing-overload-in-americas-schools* (2014).

8. Seong Jae Lee, Yun-En Liu, and Zoran Popovic. 2014. Learning Individual Behavior in an Educational Game: A Data-Driven Approach. In *Educational Data Mining 2014*.

9. Yun-En Liu, Travis Mandel, Eric Butler, Erik Andersen, Eleanor O'Rourke, Emma Brunskill, and Zoran Popovic. 2013. Predicting player moves in an educational game: A hybrid approach. In *Educational Data Mining 2013*.

10. Derek Lomas, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 89–98.

11. Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. 2014. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1077–1084.

12. Robert J Mislevy, John T Behrens, Kristen E Dicerbo, and Roy Levy. 2012. Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. *JEDM-Journal of Educational Data Mining* 4, 1 (2012), 11–48.

13. Valerie Jean Shute and Matthew Ventura. 2013. *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

14. Valerie J Shute, 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 (2009), 295–321.

15. Bulent Tastan and Gita Reese Sukthankar. 2011. Learning Policies for First Person Shooter Games Using Inverse Reinforcement Learning.. In *AIIDE*.