Analyzing Student Action Sequences and Affect While Playing Physics Playground

Juan Miguel L. Andres[1], Ma. Mercedes T. Rodrigo[1],
Ryan S. Baker[2], Luc Paquette[2], Valerie J. Shute[3], Matthew Ventura[3]

[1] Ateneo de Manila University, Quezon City, Philippines
[2] Teachers College, Columbia University, New York, NY, USA
[3] Florida State University, Tallahassee, FL, USA
{mandres, mrodrigo}@ateneo.edu,
baker2@exchange.tc.columbia.edu, luc.paquette@gmail.com,
{vshute, mventura}@fsu.edu

Abstract. Physics Playground is an educational game that supports physics learning. It accepts multiple solutions to most problems and does not impose a stepwise progression through the content. Assessing student performance in an open-ended environment such as this is therefore challenging. This study investigates the relationships between student action sequences and affect among students using Physics Playground. The study identified most frequently traversed student action sequences and investigated whether these sequences were indicative of either boredom or confusion. The study found that boredom relates to poor performance outcomes, and confusion relates to sub-optimal performance, as evidenced by the significant correlations between the respective affective states, and the student action sequences.

Keywords: Affect modeling, action sequences, boredom, confusion, Physics Playground

1 Introduction

Physics Playground (PP) is an educational game that immerses learners in a choice-rich environment for developing intuitive knowledge about simple machines. As the environment does not impose a stepwise sequence on the learner, and because some problems can have multiple solutions, learners have the freedom to explore, attempt to solve, or abort problems as they wish. The challenge these types of environments impose on educators is that of assessment. Within such an open-ended system, how do educators and researchers assess learning as well as the quality of the learning process?

This study focuses its attention on two main phenomena: student learning and student affect. Student learning within PP refers to how well a player can understand the concepts surrounding four simple machines through their efficient execution in attempting to solve levels, as evidenced by the badges they earn.

Student affect refers to experiences of feelings or emotions. In this study, the affective states of interest are confusion and boredom, as prior studies have shown them to relate significantly with learning [4, 10]. Confusion is uncertainty about what to do next [5]. Confusion is scientifically interesting because it has a positive and negative dimension, wherein it either spurs learners to exert effort deliberately and purposefully to resolve cognitive conflict, or leads learners to become frustrated or bored, and may lead to disengagement from the learning task altogether [7].

Boredom, on the other hand, is an unpleasant, transient affective state in which the individual feels a pervasive lack of interest in and difficulty concentrating on the current activity [8]. Boredom has been a topic of interest because of the negative effects usually associated with it, such as poor long-term learning outcomes when students are not provided any scaffolding [10] and its being characteristic of less successful students [11].

A study conducted by Biswas, Kinnebrew, and Segedy [2] investigated frequently traversed sequences of student actions using bottom-up, data-driven sequence mining, the results of which contributed to the development of performance- and behavior-based learner models. The analyses in this paper seek to perform similar sequence-mining methods in order to find student sequences that inform either of the affective states of interest.

This study conducted data-driven sequence-mining analyses to answer the following research questions:

1. What were the frequently traversed student action sequences among students playing Physics Playground?
2. Are these action sequences indicative of either boredom or confusion?

The analyses in this study are limited to the data collected during gameplay of Physics Playground from six data gathering sessions conducted at a public school in Quezon City in 2013. Data is limited to the interaction logs generated by the game as well as human observation of affect as logged by two coders trained in the Baker-Rodrigo-Ocumpaugh Monitoring Protocol [9].

2 Methodology

2.1 Participant Profile

Data were gathered from 60 eighth grade public school students in Quezon City, Philippines. Students ranged in age from 13 to 16. Of the participants, 31% were male and 69% were female. As of 2011, the school had 1,976 students, predominantly Filipino, and 66 teachers. Participants had an average grade on assignments of B (on a scale from A to F).

2.2 Physics Playground

Physics Playground (PP) is an open-ended learning environment for physics that was designed to help secondary school students understand qualitative physics. Qualitative physics is a nonverbal, conceptual understanding of how the physical world operates [12].

PP has 74 levels that require the player to guide a green ball to a red balloon. An example level is shown in Fig. 1. The player achieves this goal by drawing agents (ramps, pendulums, springboards, or levers) or by nudging the ball to the left or right by clicking on it. The moment the objects are drawn, they behave according to the law of gravity and Newton's 3 laws of motion [12].

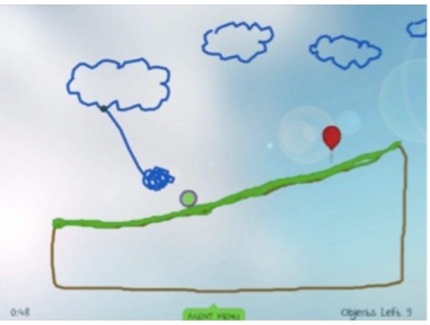

Fig. 1. Example PP level.

Performance Metrics. Gold and silver badges are awarded to students who manage to solve a level. A gold badge is given to a student who is able to solve the level by drawing a number of objects equal to the particular level's par value (i.e., the minimum number of objects needed to be drawn to solve the level). A student who solves a level using more objects will earn a silver badge. A student earns no badge if he was not able to solve the level. Many levels in PP have multiple solutions, meaning a player can solve the level using different agents.

2.3 Interaction Logs

During gameplay, PP automatically generates interaction log files. Each level a student plays creates a corresponding log file, which tracks every event that occurs as the student interacts with the game. Per level attempt, PP tracks begin and end times, the agents used, and the badges awarded upon level completion. PP also logs the *Freeform Objects* that player draw, or objects that cannot be classified as any of the four agents. The physics agents within PP are as follows:

- Ramp, any line drawn that helps to guide a ball in motion,
- Lever, an agent that rotates around a fixed point, usually called a fulcrum,
- Pendulum, an agent that directs an impulse tangent to its direction of motion,
- Springboard, an agent that stores elastic potential energy provided by a falling weight.

2.4 The Observation Protocol

The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) is a protocol for quantitative field observations of student affect and engagement-related behavior, described in detail in [9]. The affective states observed within Physics Playground in this study were engaged concentration, confusion, frustration, boredom, happiness, delight, and curiosity. The affective categories were drawn from [6].

BROMP guides observers in coding affect through different utterances, body language, and interaction with the software specific to each affective state. A total of seven affective states were coded, however, this study focuses on three: concentration, confusion, and boredom. These were identified as follows:

1. Concentration — immersion and focus on the task at hand, leaning toward the computer and attempting to solve the level, a subset of the flow experience described in [5].
2. Confusion — scratching his head, repeatedly attempting to solve the same level, statements such as "I don't understand?" and "Why didn't it work?"
3. Boredom — slouching, sitting back and looking around the classroom for prolonged periods of time, statements such as "Can we do something else?" and "This is boring!"

Following BROMP, two trained observers observed ten students per session, coding students in a round-robin manner, in 20-second intervals throughout the entire observation period of 2 hours. During each 20-second window, both BROMP observers code the current student's affect independently. If the student exhibited two or more distinct states during a 20-second observation window, the observers only coded the first state. The inter-coder reliability for affect for the two observers in the study was acceptably high with a Cohen's Kappa [3] of 0.67. The typical threshold for certifying a coder in the use of BROMP is 0.6, a standard previously used in certifying 71 coders in the use of BROMP (e.g., [9]).

The observers recorded their observations using HART, or the Human Affect Recording Tool. HART is an Android application developed to guide researchers in conducting quantitative field observations according to BROMP, and facilitate synchronization of BROMP data with educational software log data.

2.6 Data Collection Process

Before playing PP, students answered a 16-item multiple-choice pretest for 20 minutes. Students then played the game for 2 hours, during which time two trained observers used BROMP to code student affect and behavior on the HART application. A total of 4,320 observations were collected (i.e., 36 observations per participant per each of the two observers). After completing gameplay, participants answered a 16-item multiple-choice posttest for 20 minutes. The pretest and posttest were designed to assess knowledge of physics concepts, and have been used in previous studies involving PP [12].

To investigate how students interacted with PP, the study made use of the interaction logs recorded during gameplay to analyze student performance. Of the 60 participants, data from 11 students were lost because of faulty data capture and corrupted log files. Only 49 students had complete observations and logs. As a result, the analysis in this paper is limited to these students, and the 3,528 remaining affect observations. Engaged concentration was observed 72% of the time, confusion was observed 8% of the time, and boredom and frustration were observed 7% of the time. Happiness, delight, and curiosity comprise the remaining 6% of the observation time.

3 Analyses and Results

3.1 Agent Sequences

All PP-generated logs were parsed and filtered to produce a list containing only the events relevant to the study. Sequences were then separated into one of two categories: 1) silver sequences, or the sequences that ultimately led to a silver badge, which comprised 44% of all level attempts, and 2) unsolved sequences, or the sequences that led to the student quitting the level without finding a solution, which comprised 39% of all level attempts. Sequences that ended in gold badges were dropped from the analysis because they only comprised 17% of all level attempts.

Every time a student earns a badge after solving a level, the badge is awarded for one of the four agents (e.g., a player is awarded a silver ramp badge for solving the level using a ramp, and another player is awarded a gold pendulum badge for solving another level using a pendulum). We tracked the agents the badges were awarded for per level, and used this list of badges to relabel the sequences based on correctness. If the level awarded a badge for an agent, that agent was labeled as `correct` for that level; if not, the agent was labeled as `wrong` for the level. For example, on a level that awarded badges for springboards and levers, a sequence of `Lever > Ramp > Springboard > Level End (silver-springboard)` would be relabeled as `correct > wrong > correct > Level End (silver)`.

The relabeling was done because most of the sequences were level-dependent, that is, a majority of some sequences appeared on only one or two levels. By relabeling based on correctness, we were able to ensure level-independence among sequences. Sequences were tabulated and their frequencies calculated (i.e., how many times each of the 49 students traversed each of the sequences). We calculated for distribution of sequence frequencies, and the sequences we found to occur rarely (i.e., less than 30% of the population traversed them) were dropped from the analysis. We found that the gold sequences occurred rarely, which was another reason they were dropped from the analysis. The resulting silver and unsolved sequences can be found in Tables 1 and 2, respectively, along with the frequency means and standard deviations.

Table 1 lists the top 7 silver sequences within PP, which were traversed by more than 30% of the study's population. The Sequences column shows what the respective sequences look like, and the Frequency column shows the average number of times the 49 students traversed them and the standard deviations.

Highlighted sequences showed significant correlations with either boredom or confusion, as discussed further in Section 3.2. Table 2 is presented in the same manner.

Table 1. Top 7 silver sequences, their traversal frequency means, and standard deviations.

	Sequences	Frequency	
		Mean	SD
1	`correct>Level End (silver)`	3.53	2.34
2	`Level End (silver)`	2.61	2.33
3	`wrong>Level End (silver)`	1.90	1.37
4	`correct>correct>Level End (silver)`	1.61	1.15
5	`wrong>correct>Level End (silver)`	0.90	1.01
6	`correct>correct>correct>Level End (silver)`	0.80	1.00
7	`wrong>correct>correct>Level End (silver)`	0.69	0.77

The silver sequences in Table 1 show signs of experimentation, with students playing around with the correct and incorrect agents to solve the levels, as seen in sequences 5 and 7. Sequences 1, 4, and 6 show students using the correct agents, but are unable to earn gold badges. This suggests that students, while knowing which agents to use, do not have a full grasp of the physics concepts surrounding the agents' execution. Sequence 3 shows students using wrong objects to solve the levels. While this may suggest that students are still struggling to understand how the agents work and which agent would best solve a level given the ball and the balloon's positions, this may have also been caused by the PP logger labeling the objects they drew as freeform objects, and not one of the correct agents.

Sequence 1 shows the students drawing only the correct agent, but are still unable to earn a gold badge. The sequence-mining algorithm only pulled events related to drawing any of the four main agents, which are enumerated in Section 2.3. Drawing a lever or a springboard, for example, would require drawing more than one component. A lever requires the fulcrum, the board, and the object dropped on the board to project the ball upwards. In order for the agent to work, it has to be executed correctly (i.e., the board must be long enough, with the fulcrum in the right position, and the object dropped on the board must be heavy enough to propel the ball into the air). Sequence 1 may have been caused by students drawing the correct agent, but improperly executing it. For example, the student may not have drawn the right-sized weight to drop on the lever, and thus had to draw another. While drawing another weight to drop on the lever counts towards the level's object count, it was not logged as a separate event by the sequence mining analysis because the player did not draw another agent, only a component of it. Sequence 2, on the other hand, is suspect because despite the student drawing no objects to solve a level, he ends up with only a silver badge. This was most likely caused by the improper logging of the game. The top 7 most frequently traversed silver sequences account for 58% of the total number of silver sequences.

Table 2. Top 6 unsolved sequences, their traversal frequency means, and standard deviations.

	Sequences	Frequency	
		Mean	SD
1	`Level End (none)`	10.69	8.17
2	`wrong>Level End (none)`	1.55	1.65
3	`correct>Level End (none)`	1.29	1.50
4	`wrong>wrong>Level End (none)`	0.45	0.65
5	`correct>correct>Level End (none)`	0.41	0.73
6	`wrong>correct>Level End (none)`	0.39	0.57

Table 2, which shows the top 6 unsolved sequences, shows signs of students giving up. Sequence 1 shows students giving up without even drawing a single object, which could have been caused by one of two things: 1) the student saw the level and decided to quit without attempting to solve it, or 2) again, the logger did not log the objects correctly. This sequence is similar to one of the silver sequences in that no objects were drawn. What makes them different, however, is what the sequences ultimately led to. The silver sequences ended in a silver badge, and the unsolved sequences ended in the student earning no badge. The majority of the sequences listed in Table 2 show students experimenting mainly with wrong objects, whether agents or freeform objects. This implies that the students are lacking in the understanding of how to solve the levels. Sequences 3 and 5 are interesting because it is unclear whether or not the students understood the concepts of the agents. That is, students were drawing the correct agents, but could not get the ball to reach the balloon. Despite drawing one or two correct agents, the students decided to give up and quit. The top 6 unsolved sequences account for 81% of the total number of unsolved sequences.

3.2 Relationship with Affect

We computed frequencies for each of the 13 sequences that the 49 students traversed. Correlations were then run between each of the 13 arrays and the incidences of confusion and boredom. Because the number of tests introduces the possibility of false discoveries, Storey's adjustment [13] was used as a post-hoc control, which provides a q-value, representing the probability that the finding was a false discovery. Tables 3 and 4 show the results. Highlights and asterisks (*) were used on significant findings ($q \leq 0.05$).

Table 3 lists the top 7 most frequently traversed silver sequences, from left to right. The sequences these header numbers represent can be found in Table 1. The table shows the correlation between each of the top 7 silver sequences using a metric that represents the percentage of all attempts that match each of the sequences, the percentage of time the students were observed to be confused (r, con), and the percentage of time the students were observed to be bored (r, bor).

Table 4 is presented in the same manner, with sequence information in Table 2 for the top 6 unsolved sequences.

Table 3. Correlations between top 7 silver sequences, confusion, and boredom.

	Top 7 silver sequences						
	1	2	3	4	5	6	7
r, con	-0.33	0.23	0.41*	0.03	0.17	0.54*	0.28
r, bor	-0.20	-0.17	-0.19	-0.05	0.14	-0.19	-0.20

Table 3 shows two significant positive correlations between confusion and the silver sequences. The two sequences showed signs of lesser understanding of the agents. Sequence 3 shows students using only a wrong object to solve a level, which may have been caused either by incorrect object labeling (e.g., PP logged a ramp as a Freeform Object), or the student found a different way of solving the level. Like in most learning environments, players are able to game the system – or systematically misuse the game's features to solve a level [1] – within PP through stacking. Stacking is done when players draw freeform objects to either prop the ball forward or upward, which may have been the case in sequence 3. Sequence 6 shows students drawing only correct agents. These sequences having significant correlations with confusion may imply lesser understanding among confused students as the they are not only dealing with proper agent execution, but also with deciding which agent would best solve the level. Despite the challenges faced by these students, however, they still managed to find a solution to the level. Our findings suggest that the inability to grasp the physics concepts surrounding the agents is a sign of confusion.

Table 4. Correlations between top 6 unsolved sequences, confusion, and boredom.

	Top 6 unsolved learning sequences					
	1	2	3	4	5	6
r, con	-0.17	0.00	-0.12	-0.01	-0.06	0.04
r, bor	-0.12	0.13	0.12	-0.03	0.48*	0.06

Table 4 shows that one of the most frequently traversed unsolved sequences has a significant positive correlation with boredom. This sequence shows students using only correct agents, but ultimately deciding to give up. This may have been caused by the inability to execute the agents correctly, which may imply that, unlike confused students, bored students were not likely to exert additional effort to try to solve the level or understand proper agent execution. As mentioned previously, boredom has been found to have significant relationships with negative performance outcomes. In this case, sequences all ultimately led to disengagement: students quitting the level before finding a solution, showing signs of giving up and lack of understanding of any of the four agents.

4 Conclusions and Future Work

This study sought to identify the most frequently traversed student action sequences among eighth grade students while interacting with an education game for physics called Physics Playground. Further, the study sought to investigate how these sequences may be indicative of affective states, particularly boredom and confusion, which have been found to significantly affect student learning.

Data-driven sequence mining techniques were conducted to identify most frequently traversed actions sequences in two categories: the sequences that would eventually lead the student to a silver badge, and the paths that would eventually lead the student to not earning a badge.

In the silver sequences, students played around with freeform objects and some of the four agents in attempting to solve the level. The study found confusion to correlate significantly with two of the silver sequences, which supports previous findings regarding the relationship between confusion and in-game achievement, which suggest that because students are unable to grasp the concepts surrounding the agents and their executions, students resort to finding other solutions.

In the unsolved sequences, students would give up and quit without finding a solution, despite already using the correct agents to solve the level. The study found boredom to correlate significantly with one of the unsolved sequences. This finding supports the literature that has shown that boredom relates to poor learning outcomes. This work provides further evidence that boredom and disengagement from learning go hand-in-hand.

This study provides specific sequences of student actions that are indicative of the boredom and confusion, which has implications on the design and further development of Physics Playground. This study also contributes to the literature by providing empirical support that boredom and confusion are affective states that influence performance outcomes within open-ended learning environments, and are thus affective states that learning environments must focus on detecting and providing remediation to. We found that both bored and confused students will tend to continuously use correct agents in attempting to solve levels, but execute them incorrectly. The difference between the two, however, is that confused students tend to end up solving the level, while bored students give up.

The analyses run in this paper were part of a bigger investigation, and as such, there are several interesting ways forward in light of our findings. The paper aims for its findings to contribute to the creation of a tool that can automatically detect affect given a sequence of student interactions, and provide necessary remediation in order to curb student experiences of boredom.

Relationship analyses run between student action sequences and incidences of affect in this paper were done through correlations. However, findings were not able to determine whether boredom or confusion occurred more frequently during specific action sequences. We want to find out whether boredom or confusion occurred before, during, or after the students' execution of the action sequences, and in doing so, see whether or not the affective states were causes or effects of the action sequence executions. We are currently investigating this relationship in a separate study.

Acknowledgements. We would like to thank the Ateneo Center for Educational Development, Carmela C. Oracion, Christopher Ryan Adalem, and the officials at Krus Na Ligas High School, Jessica O. Sugay, Dr. Matthew Small, and the Gates Foundation Grant #OP106038 for collaborating with us.

References

1. Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 383-390). ACM.
2. Biswas, G., Kinnebrew, J. S., & Segedy, J. R. (2011). Using a Cognitive/Metacognitive Task Model to analyze Students Learning Behaviors.
3. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20 (1960), 37-46.
4. Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. Journal of Educational Media, 29(3), 241-250.
5. Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New Y ork: Harper Perennial.
6. D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2005). Integrating affect sensors in an intelligent tutoring system. In Proceedings of the Workshop on Affective Interactions: The computer in the affective loop workshop, International conference on intelligent user interfaces (pp. 7- 13). New York: Association for Computing Machinery.
7. D'Mello, S., Graesser, A. (2012). Dynamics of affective states during complex learning. Learning and Instruction, 22(2): 145-157.
8. Fisherl, C. D. (1993). Boredom at work: A neglected concept. Human Relations, 46(3), 395-417.
9. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
10. Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, 1(1), 107-128.
11. San Pedro, M. O. Z., d Baker, R. S., Gowda, S. M., & Heffernan, N. T. (2013, January). Towards an understanding of affect and knowledge from student interaction with an Intelligent Tutoring System. In Artificial Intelligence in Education (pp. 41-50). Springer Berlin Heidelberg.
12. Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. The Journal of Educational Research, 106(6), 423-430.
13. Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64: 479-498.