Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

An Exploratory Analysis of Confusion Among Students Using Newton's Playground

Juan Miguel L. ANDRES [a*], Ma. Mercedes T. RODRIGO[a], Jessica O. SUGAY[a], Ryan S. BAKER[b], Luc PAQUETTE[b], Valerie J. SHUTE[c], Matthew VENTURA[c], & Matthew SMALL[c]

[a]*Ateneo de Manila University, Philippines*
[b]*Columbia University, USA*
[c]*Florida State University, USA*
miglimjapandres@gmail.com, mrodrigo@ateneo.edu, jsugay@ateneo.edu
baker2@exchange.tc.columbia.edu, luc.paquette@gmail.com
vshute@fsu.edu, mventura@fsu.edu, matthewsmall@gmail.com

Abstract: We investigated the interplay between confusion and in-game behavior among students using Newton's Playground (NP), a computer game for physics. We gathered data from 48 public high school students in the Philippines. Upon analyzing quantitative field observations and interaction logs generated by NP, we found that confusion among students was negatively correlated with earning a gold badge (solving a problem with objects under par), positively correlated with earning a silver badge (solving a problem with objects over par), and positively correlated with stacking (drawing numerous small objects to reach the objective), a form of gaming the system.

Keywords: Confusion, student affect, Newton's Playground, learning, in-game behavior

1. Introduction

In recent years, researchers have been investigating the state of confusion among learners using intelligent tutoring systems. Confusion, or cognitive disequilibrium, is defined as the uncertainty about what to do next (D'Mello et al., 2005). It occurs when a student encounters stimuli or experiences that fail to meet expectation (D'Mello, Lehman, Pekrun, & Graesser, 2014) and plays an important role in the learning process because cognitive disequilibrium "has a high likelihood of activating conscious, effortful cognitive deliberation, questions and inquiry that aim to restore cognitive equilibrium (Craig, Graesser, Sullins, & Gholson, 2004; D'Mello et al., 2008)." Confusion is actually useful when it spurs learners to exert effort deliberately and purposefully to resolve cognitive conflict. If the learners are successful, they return to a state of flow – complete immersion and focus upon the system (Csikszentmihalyi, 1990). However, confusion may also be negative. If unresolved, confusion can lead to frustration or boredom, and students may decide to disengage from the learning task altogether (D'Mello & Graesser, 2012). Previous studies have shown confusion to correlate positively with learning gains (Craig et al., 2004; D'Mello et al., 2014). Extended periods of confusion, however, were associated with negative learning outcomes (Lee, Rodrigo, Baker, Sugay, Coronel, 2011).

This study explores student in-game events that may be indicative of student confusion within Newton's Playground (NP, described in detail in Section 2), a computer game for physics. NP requires the player to guide a green ball to a red balloon by drawing simple machines on the screen with colored markers controlled by the mouse. The software has been used previously for stealth assessment of creativity, persistence, and conceptual physics understanding (Shute & Ventura, 2013). The studies found significant improvement in terms of conceptual physics understanding among students that played the game, depending on how engaged they were (or how many levels they attempted to play) during gameplay (see Shute, Ventura, & Kim, 2013). Additional studies are using Newton's Playground to examine the relationship between student affect, in-game behavior, and

mastery of physics concepts. This exploratory study examines human observations alongside logged student-software interactions to determine what in-game events correlate with student confusion.

2. Study population, system, and data collection methodology

2.1 Participant Profile

We conducted a study to measure the relationship between a variety of affective and cognitive variables. Data was gathered from 60 eighth grade public school students in Quezon City, Philippines. Students ranged in age from 13 to 16. As of 2011, the school had 1,976 students, predominantly Filipino, and 66 teachers. Of the participants, 31% were male and 69% were female. Participants were asked to rate how frequently they played video games and watched television on a scale of 1 (not at all) to 7 (everyday, for more than 3 hours), and the resulting average frequency of gameplay is 3.2 (in between a few times a month, and a few times a week), and the resulting average frequency of watching television is 5.9 (in between everyday, but for less than 1 hour, and everyday, for 1-3 hours). Participants were asked for their most frequent grade on assignments, and on a scale of 0 (F) to 4 (A), the average most frequent grade of the participants is 3.1 (B).

2.2 The Software

Newton's Playground (NP) is a computer game for physics patterned after Crayon Physics Deluxe. It was designed to help secondary school students understand qualitative physics (Shute & Ventura, 2013). Qualitative physics is a nonverbal conceptual understanding of how the physical world operates, along the lines of Newtonian physics. Qualitative physics is characterized by an implicit understanding of Newton's three laws: balance, mass, and conservation and transfer of momentum, gravity, and potential and kinetic energy (Shute et al., 2013).

NP is a two-dimensional computer-based game that requires the player to guide a green ball to a red balloon. Two example levels are shown in Figure 1, the level on the left requiring a pendulum, and the level on the right requiring a lever. The player uses the mouse to nudge the ball to the left and right (if the surface is flat), but the primary way to move the ball is by drawing or creating simple machines on the screen with the mouse and colored markers. The objects come to life once the object is drawn. Everything obeys the basic rules of physics relating to gravity and Newton's three laws of motion (Shute et al., 2013).

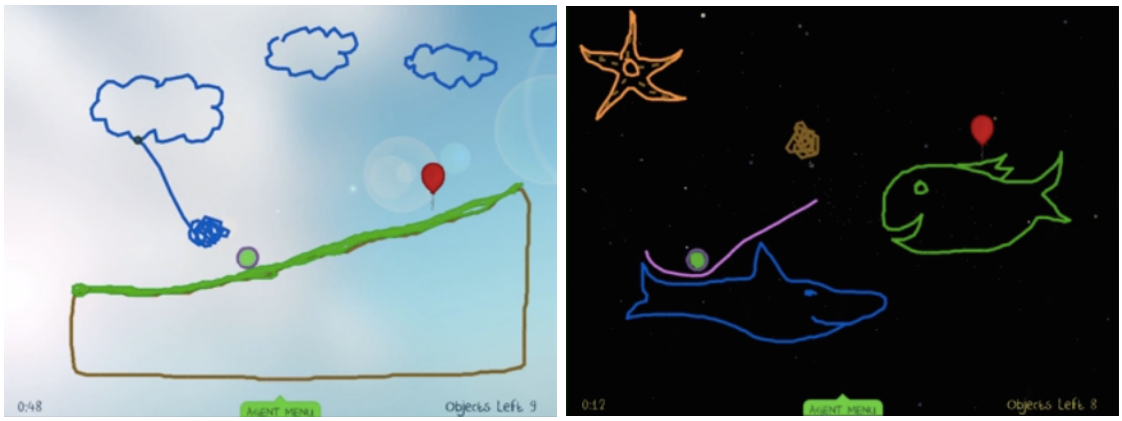

Figure 1. Examples of Newton's Playground levels.

The 74 levels in NP require the player to solve the problems via drawing different simple machines, representing agents of force and motion: inclined plane/ramps, levers, pendulums, and springboards. Again, all solutions are drawn with colored markers using the mouse. A ramp is any line drawn that helps to guide a ball in motion. A ramp is useful when a ball must travel over a hole. A lever rotates around a fixed point, usually called a fulcrum or pivot point. Levers are useful when a player wants to move the ball vertically. A swinging pendulum directs an impulse tangent to its

direction of motion. The pendulum is useful when the player wants to exert a horizontal force. A springboard (or diving board) stores elastic potential energy provided by a falling weight. Springboards are useful when the player wants to move the ball vertically.

Gold badges versus silver badges. Some levels in NP have multiple solutions, which means a player can solve the level using different agents. Gold badges are awarded when a player solves a problem "under par;" that is, under a limit set for a specific solution. For example, a level may be solved using a ramp, with a par of 1 object, or a lever, with a par of 3 objects. If a player solves the level with more objects than par, he receives a silver badge. Gold badges suggest that the player has mastered the agent relevant to the given level. Silver means the player may not have fully mastered the agent yet.

Stacking. During pilot testing, Shute et al. (2013) reported that it was possible to game the system to succeed without using the target knowledge (Baker, Corbett, Koedinger, & Roll, 2006) by drawing many tiny objects that stack up, propelling the ball upward until it reaches the target. This behavior is called stacking. The log files capture stacking actions on levels where the player did this (Shute et al., 2013).

2.3 The Interaction Logs

We collected two types of data during the study: interaction logs and human observations. During gameplay, NP automatically generates log files. Each level a student plays creates a corresponding log file, which tracks every interaction the student has with the game in terms of particular counts and times for selected features of gameplay. These features include but are not limited to:

- Time spent on the level in seconds,
- Number of in-level restarts,
- Number of objects drawn in a solution attempt,
- Whether the level was ultimately solved,
- Whether or not the player earned a gold or silver badge, and
- Whether or not a player was stacking, a form of gaming the system – the systematic misuse of system features to advance through the learning materials without learning the content (Baker et al., 2006) – within Newton's Playground

Each of these variables provides useful information about students' gameplay behaviors, which can then be used to make inferences about how well they are doing in the game (Shute et al., 2013).

2.4 The Observation Protocol

The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) is a protocol for quantitative field observations of student affect and behavior. BROMP is a holistic coding procedure that has been used in thousands of hours of field observations of students, from kindergarten to undergraduate populations. It has been used for several purposes, including to study the engagement of students participating in a range of classroom activities (both activities involving technology and more traditional classroom activities) and to obtain data for use in developing automated models of student engagement with Educational Data Mining (EDM) (Ocumpaugh, Baker, & Rodrigo, 2012). Within BROMP, each student observation lasts 20 seconds, and the observers move from one student to the next in a round robin manner during the observation period.

The affective states observed within Newton's Playground were concentration, confusion, frustration, boredom, happiness, delight, and curiosity. The behaviors observed were on-task, off-task, stacking, and a behavior called *without thinking fastidiously* (WTF), a behavior in which, despite a student's interaction with the software, "their actions appear to have no relationship to the intended learning task (Wixon, Baker, Gobert, Ocumpaugh, & Bachmann, 2013)." The analysis of the behaviors, however, is outside this paper's scope.

The inter-coder reliability for affect was acceptably high with a Cohen's (1960) Kappa of 0.67. The typical threshold for certifying a coder in the use of BROMP is 0.6, established across dozens of studies as well as the previous affective computing literature.

2.5 Procedure

Before playing Newton's Playground, students completed a 16-item multiple-choice pretest for 20 minutes. Students were then assigned a computer on which they would play NP. Students played the game for two hours, during which, two trained observers used BROMP to code student affect and behavior. A total of 36 observations per participant per observer were collected. Videos of participants' faces were also recorded during gameplay. After completing the two hours of gameplay, participants completed a 16-point multiple-choice posttest for 20 minutes. The pretest and posttest were designed to assess knowledge of physics concepts, and has been used in previous studies involving Newton's Playground (Shute et al., 2013).

The pretest and posttest scores were tabulated and averaged. Students scored an average of 6.02 out of 16 in the pretest, and 6.02 out of 16 in the posttest. While these results suggest that NP did not seem to help increase knowledge of physics concepts, the researchers noticed that students were answering the posttest hurriedly. The posttest scores may thus not reflect an accurate knowledge assessment. It is important to note, however, that significant pretest-to-posttest improvements were reported in three previous studies that also used NP. Students in these studies used NP for longer periods of time.

3. The Relationship between Student Confusion and In-game Events

In order to investigate how students mastered content in Newton's Playground, we made use of the interaction logs recorded during gameplay to analyze student performance. Of the 60 participants, data from 12 students were lost because of faulty data capture and corrupted log files. Only 48 students had complete observations and logs. The analysis that follows is limited to these students.

The BROMP observations were tabulated, and the percentage of each affective state per student was calculated. Boredom, confusion, and frustration were three of the more commonly observed affective states, besides concentration.

All interaction logs were passed through a parser to arrange log events in tab-delimited text files. These text files were then run through a filter to get per student, per level, per attempt summaries, such as total time spent, total number of restarts, total number of objects drawn, etc. Finally, the information was collapsed to form per student vectors that summarized the students' entire interactions with the game. Each vector included the following attributes, which are indicative of mastery:

- Gold badge – percentage of level attempts solved, earning the student a gold badge
- Silver badge – percentage of level attempts solved, earning the student a silver badge
- Stacking – percentage of level attempts wherein a student was flagged for stacking

These three attributes, among about thirty other gameplay features, were correlated with the percentage of confusion, based on the human observations. Because the number of tests introduces the possibility of false discoveries, Storey's adjustment was used as a post-hoc control. Storey's translates the p-value to a q-value, which represents the probability that the finding was a false discovery. Among the results, earning a silver badge was positively correlated with confusion, while earning a gold badge was negatively correlated, and stacking was positively correlated with confusion. Table 1 shows their correlations, p-values, and q-values. Note that the findings were still significant even after the post-hoc correction was applied.

Table 1: Correlations between student interaction and confusion, their p-values, and q-values.

	Silver badges	Gold badges	Stacking
Correlation	0.32	-0.29	0.31
p-value	0.03	0.04	0.03
q-value	0.02	0.03	0.03

The opposite relationship between gold and silver badges and confusion is interesting. Recall that solving a level under par earns a player a gold badge, while solving a level with more objects than par, no matter how many are drawn, earns the student a silver badge. As mentioned earlier, earning a

gold badge is indicative of mastery of the four agents being used in the game. Mastery of these agents goes beyond knowing what agents to use, as in many cases, different agents can be used to solve the same level. Mastery also entails proper execution of the drawing, and being able to keep the number of objects under par. A player has to be very precise in his understanding of various aspects of the agents (e.g. how massive an object must be and from what height it must be dropped onto a springboard in order to propel the ball towards the target) Students who have mastered the agents do not have to experiment for prolonged periods of time with drawing different objects to see which agents will propel the ball closer to the balloon.

If a student has not yet mastered the agents, however, he may end up making more guesses, experimenting by drawing different objects until the ball reaches the balloon, and ends up earning a silver badge. Understanding the solution could help the student gain mastery of the agents, but finding a solution merely by chance may contribute to making him even more confused.

The other interesting observation was the positive correlation between confusion and stacking, which, as mentioned earlier, is a form of gaming the system within NP, a behavior associated with negative affect. Previous studies have found confusion to have no significant effect on gaming the system (Baker, D'Mello, Rodrigo, & Graesser, 2010). This correlation, however, suggests that the more confused a student is, the more likely he is to stack. Stacking indicates a lack of mastery of the physics agents.

4. Conclusions and Future Work

In this study, we attempted to identify in-game events that may relate to confusion among students playing Newton's Playground. Students played NP for two hours while two BROMP coders labeled student affect and behavior. These observations were then analyzed alongside NP interaction logs. In our analysis, we found that confusion was negatively correlated with earning a gold badge but positively correlated with earning a silver badge, and that stacking and confusion are positively correlated. This implies that, within our population, students who are confused lack of mastery of physics concepts. They solve problems inefficiently, using more objects than necessary. On the other hand, students who develop mastery of physics concepts are able to solve the NP problems with an optimal number of objects.

This study is the first of many analyses done on the data set, and is part of a bigger investigation. As such, there are several next steps to take from this work. One avenue is to disambiguate good and bad confusion, and find what student behaviors are indicative of each. Learning benefits can be derived from episodes of good confusion (D'Mello & Graesser, 2011). Bad confusion, on the other hand, has no pedagogical value (D'Mello & Graesser, 2011).

It would also be interesting to explore how student boredom manifests itself within NP. Boredom is defined as an "unpleasant, transient affective state in which the individual feels a pervasive lack of interest in and difficulty concentrating on the current activity" (Fisher, 1993). Boredom has been associated with poorer learning (e.g. Craig et al., 2004) and problem behaviors, such as gaming the system (e.g. Baker et al., 2010).

Acknowledgements

We would like to thank the Ateneo Center for Educational Development, Carmela C. Oracion, Christopher Ryan Adalem, the officials at Krus Na Ligas High School, and the Gates Foundation Grant #OP106038 for making this study possible.

References

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Roll, I. (2006) Generalizing Detection of Gaming the System Across a Tutoring Curriculum. Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 402-411.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three

different computer-based learning environments. International Journal of Human-Computer Studies, 68(4), 223-241.
Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20 (1960), 37-46.
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.
Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New York: Harper Perennial.
D'Mello, S., Craig, S. D., Gholson, B., Franklin, S., Picard, R., & Graesser, A. C. (2005). Integrating affect sensors in an intelligent tutoring system. In Affective Interactions: The Computer in the Affective Loop Workshop at (pp. 7-13).
D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. Cognition & Emotion, 25(7), 1299-1308.
D'Mello, S., Graesser, A. (2012). Dynamics of affective states during complex learning. Learning and Instruction, 22(2): 145-157.
D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., ... & Graesser, A. (2008). AutoTutor detects and responds to learners affective and cognitive states. In Proceedings of the Workshop on Emotional and Cognitive issues in ITS in conjunction with the 9th International Conference on Intelligent Tutoring Systems (pp. 31-43).
D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. Learning and Instruction, 29, 153-170.
Fisher, C. D. (1993). Boredom at work: A neglected concept. Human Relations, 46(3), 395-417.
Lee, D.M., Rodrigo, M.M., Baker, R.S.J.d., Sugay, J., Coronel, A. (2011) Exploring the Relationship Between Novice Programmer Confusion and Achievement. Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction.
Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. (2012) Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
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.
Shute, V., & Ventura, M. (2013). Stealth assessment: Measuring and supporting learning in video games. MIT Press.
Wixon, M., Baker, R. S., Gobert, J. D., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In User Modeling, Adaptation, and Personalization (pp. 286-296). Springer Berlin Heidelberg.