Learning and Affect Trajectories Within Newton's Playground

Juan Miguel L. ANDRES[a*] **& Ma. Mercedes T. RODRIGO**[a]
[a]*Department of Information Systems and Computer Science,*
Ateneo de Manila University, Philippines
*miglimjapandres@gmail.com, mrodrigo@ateneo.edu

Abstract: Learning trajectories are typical, predictable sequences of thinking that emerge as students develop understanding of an idea. They have been used principally for research on instructional decision-making, but have also played a significant role in conducting research on learning. Affect refers to experiences of feelings or emotions. The affective states of boredom and confusion, in particular, have been of interest to researchers due to their significant relationships with student learning. In an attempt to account for gains or lack thereof among users of educational software, this study investigates these areas to monitor how students in the Philippines use software or how they are feeling compared to parallel studies conducted in the United States. In particular, this study investigates the relationships between learning trajectories and affect among students in the Philippines using Newton's Playground, a learning game for physics.

Keywords: Learning trajectories, affect trajectories, Newton's Playground

1. Theoretical Framework: Learning Trajectories and Affect

Learning trajectories (LTs) represent the "paths by which learning might proceed (Simon, 1995)." They are "typical, predictable sequences of thinking that emerge as students develop understanding of an idea (Daro, Mosher, & Corcoran, 2011)." LTs have been a topic of interest in recent years, enabling researchers to gain a better understanding of student learning. The study of LTs is less then two decades old (Clements & Sarama, 2004), and has only recently been getting attention in the field of learning sciences. Studies show that as researchers and teachers make sense of learning trajectories, they in turn can support growth in knowledge and further student learning.

Studying learning in terms of affective factors has also been a topic of interest in recent years. Two affective states of interest to researchers are confusion and boredom. Confusion or cognitive disequilibrium is the uncertainty about what to do next (D'Mello, Craig, Gholson, Franklin, Picard, & Graesser, 2005). It is interesting because it has a positive and negative dimension (D'Mello & Graesser, 2012), 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 (D'Mello & Graesser, 2012).

Boredom, on the other hand, is defined by Fisherl (1993) as an "unpleasant, transient affective state in which the individual feels a pervasive lack of interest in and difficulty concentrating on the current activity." It has been a topic of interest because of the negative effects usually associated with it, like poorer learning and problematic behaviors, such as gaming the system (Baker, D'Mello, Rodrigo, Graesser, 2010; Rodrigo, Baker, & Nabos, 2010).

The extent to which students learn from educational software is influenced by the effectiveness with which they use software and how they feel while using the software. However, we do not always monitor how students use software or how they are feeling, so we cannot always account for gains or lack of gains. This study is an in-depth examination of software usage and affect and their interactions. Specifically, this study seeks to investigate the relationships between learning trajectories and the affective states of boredom and confusion.

2. 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 Newton's Playground

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, & Kim, 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 by drawing simple machines on the screen with colored markers controlled by the mouse. An example level is shown in Figure 3.1. 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).

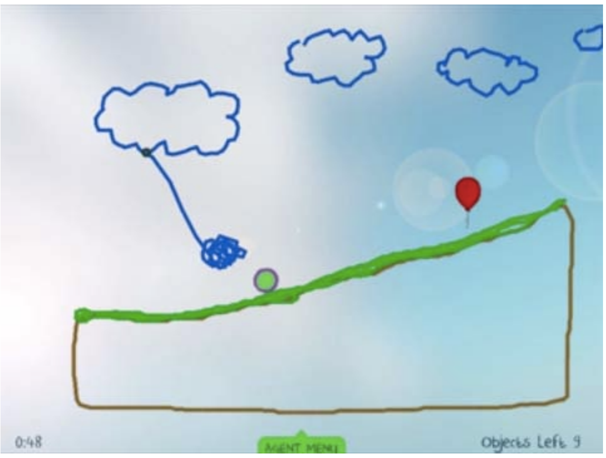

Figure 1. Example level of Newton's Playground.

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 pendulum, 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.

2.3 NP Interaction Logs

We collected two types of data during the study: interaction logs and human observations. During gameplay, NP 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. The events per level and their respective attributes that are relevant to this study are:

- Level Start,
- Level Restart,
- Level End, an event that signals the player solved the level,
    - Badge, an attribute that states the type of badge (i.e. gold or silver) awarded to the player
    - Agent, an attribute that states for which agent the badge was awarded for
- Menu Focus, an event that signals the player gave up and quit the level without solving it,
- Drawing of any of the four agents,
- Object Limit, an event that is triggered by the player reaching the maximum number of objects drawn, and
- Stacking, an event that signals the player is gaming the system.

Each of these features 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 Baker-Rodrigo-Ocumpaugh Monitoring 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, 2012)."

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 The Human Affect Recording Tool

The Human Affect Recording Tool, or HART, is an Android application developed to guide researchers in conducting quantitative field observations according to the BROMP protocol. The application synchronizes the coded observations to Internet time, allowing for more precise synchronization with log file data from the educational software under study.

HART asks for input regarding school and classroom information, coding schemes to be used, and the student IDs of the students to be observed during the session. The application then presents the student IDs in the order entered, allowing BROMP observers to more conveniently code affect and behavior until the session is manually terminated. All observations are logged on a text file that is locally stored on the device used to run HART. The application and all its functions are discussed in more detail in (Ocumpaugh et al., 2012).

2.6 Data Gathering Process

Before playing NP, 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 NP (Shute et al., 2013).

In order to investigate learning within 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 neatly 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. All statistical analyses conducted within this study are limited to the computation of percentages and result visualization.

3. Findings

We collected pre-test and post-test data from each student (N=60). Scores were generally poor. Students averaged 6.02 correct answers on both the pre-test and the post-test, out of a highest possible score of 16. This indicates that no learning improvements were detected. What follows is a descriptive analysis of the gathered data using methods described in the following subsections.

We operationalize learning trajectories on two levels:

1. On a coarse-grained level, learning trajectories are the performances of students in terms of gold and silver badges earned during gameplay, and
2. On a fine-grained level, learning trajectories are the sequences of students' interaction behaviors while solving or not solving a level.

As mentioned previously, affect coders followed BROMP, which resulted in 36 observations per student, per observer. For the purposes of this study, we define the incidence of affect as the percentage of students observed to be in a specific affective state during one observation count. We operationalize affect trajectories as the incidence of affect over time, that is, over the span of the 36 observations.

The findings in this section are from analyses conducted in finding:

1. The players' coarse-grained learning trajectories within NP,
2. The players' boredom and confusion trajectories, and
3. The relationships mined between the two.

3.1 Coarse-Grained Learning Trajectory Analysis

For the coarse-grained LT analysis, the percentages of students earning gold, silver, or no trophies were graphed over their opportunities to practice each of the four agents.

The three performance metrics (i.e. earning a gold trophy, earning a silver trophy, and earning no trophy) were used to track how well a student performed during gameplay, and in turn, see how well they understood each of the four agents used in the game. Every time a badge is awarded to a student, it is awarded for a specific agent. If a ramp was used to earn a gold badge, the student will get a gold badge for the ramp for that level. This is especially important for levels wherein any of the agents can be used to solve a level. Most levels, as the data showed however, only award badges for one of the four agents.

Using the logs generated by NP, trophies were grouped by level and by agent. In doing so, we were able to track which agents were awarded medals per levels, thus determining which agents were needed to solve each of the levels. Table 1 shows the tally of the first ten levels.

Table 1: Tally of trophies, by level and by agent for the first ten levels.

Level	Ramp	Lever	Pendulum	Springboard
P01L01	60			
P01L02	47			
P01L03		51		
P01L04			48	
P01L05				51
P02L01	47			
P02L02			46	
P02L03	46			
P02L04		46		
P02L05		22	46	30

This table shows which agents levels awarded badges for, whether gold or silver, which in turn gives us an idea of what agents were intended to solve the levels. The solution to Playground 1, Level 1, for example, is a ramp, with all sixty attempts on the level being awarded with ramp badges.

Using this table, we selected the first ten levels in which each of the four agents was used as a solution. Each of these first ten levels was then treated as an opportunity to practice one of the four agents. An opportunity to practice refers to a chance given to the student to exercise a specific skill, e.g. constructing a pendulum that pushes the ball to the target. Every level is solvable using one or more of these agents; therefore every level has one or more opportunities to practice possibly a variety of skills.

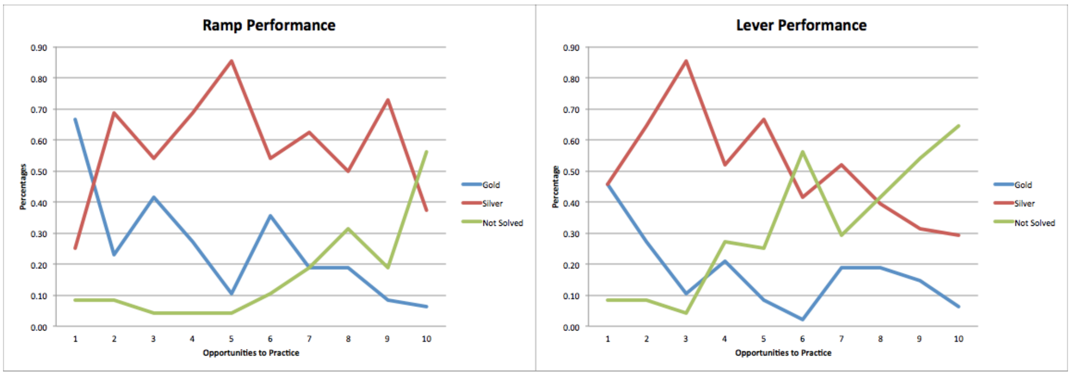

Figure 2. Learning trajectories of ramp and lever levels.

It is important to note that students were free to choose the levels they wanted to solve. The software did not force them through the material in a stepwise fashion. Furthermore, levels were not grouped thematically, by agent, so even if a student solved the levels sequentially, he would have opportunities to practice different agents.

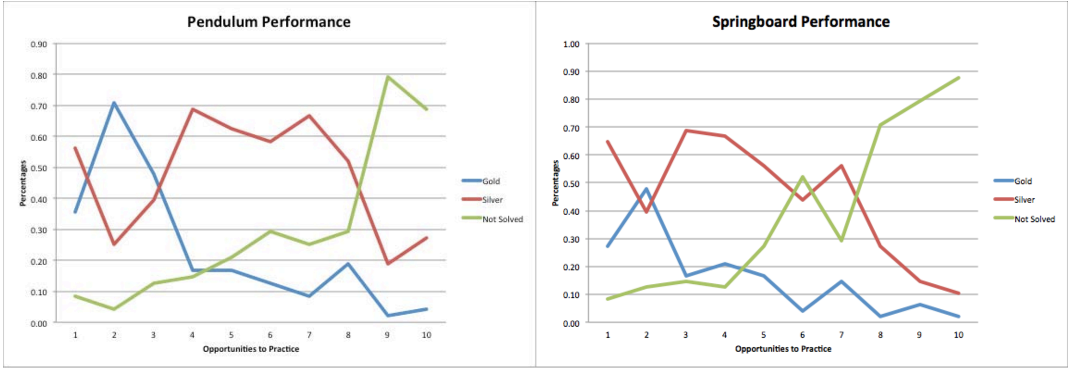

Figure 3. Learning trajectories of pendulum and springboard levels.

The percentage of students that earned gold trophies, silver trophies, and no trophies for each of the opportunities was then graphed. Figures 2 and 3 show the students' learning trajectories for each of the four agents. However, these graphs need to be unpacked further because students could choose what levels to solve, the actual levels corresponding to opportunities to practice 1 through 10 varied per student. That is for example, student 1's first opportunity to practice may be different from student 2's.

A consistent pattern can be observed across all four graphs, that is, as students progress through each of their ten opportunities to practice each of the four agents within NP, the perform more poorly over time. The percentage of students earning gold and silver trophies decreases over time, while the percentage of students unable to solve levels increases.

3.2 Fine-Grained Learning Trajectories

For the fine-grained LT analysis, a sequence mining analysis is to be conducted, taking into consideration the common sequences of actions student took in trying to solve each level.

As previously mentioned, NP generated interaction log files per level attempt that track every event that occurs as the player tries to solve the level. A filter was developed to pull only the relevant events from the log files. This filter is to be run on each level, arranging events chronologically on an output text file, divided by student. Events are to then be placed on a previous state-current state table to track transitions and transition frequencies between states. Using frequency calculations, common paths can easily be graphed and tracked through interaction network diagrams. This analysis is still in progress.

3.3 Affect Trajectory Analysis

As seen in the results above, students performed more poorly as they progressed through the levels in the game, and that at a certain point, the number of students earning trophies would just continuously decrease. The hypothesis this analysis sought to prove was that affect experienced by the students during gameplay could somehow be related to the students' eventual poor performance.

Using the logs generated by HART, all human observations per student were lined up on an Excel sheet. Each observer had a total of 36 observations per student. Using both of the observers' logs, a total of 72 observations per student were recorded. The percentage of students who were observed to be bored and the percentage of students who were observed to be confused per observation were calculated. An average percentage between the two coders was then calculated for each of the 36 observations. Figure 4 shows the affect trajectories of both confusion and boredom over time.

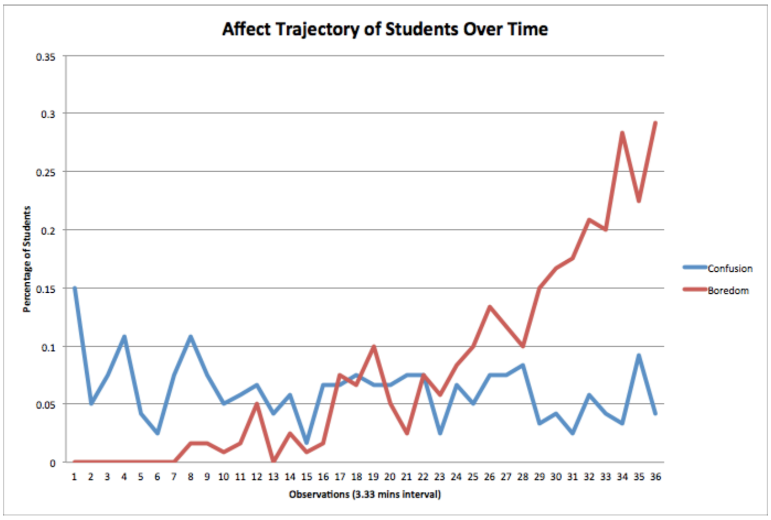

Figure 4. Affect trajectories for boredom and confusion.

It is interesting to note that while confusion was experienced by a steady number of students during the entire 2-hours of gameplay, the number of students experiencing boredom increased as the session progressed. The increase in percentage of boredom begins at observation 21, which is about one hour and ten minutes into the session.

3.4 Relationships

This study hopes to examine two relationships:

1. Between the coarse-grained LTs and affect, finding relationships between overall in-game student performance and incidences of both boredom and confusion, and
2. Between the fine-grained LTs and affect, finding relationships between common sequences and incidences of both boredom and confusion.

3.4.1 Coarse-Grained LTs and Affect

The BROMP observations were tabulated, and the percentage of each affective state per student was calculated. 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 gameplay features, 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 performance metrics:

- 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

These two metrics were correlated with the students' respective percentages of boredom. The analysis, however, found no significant relationships. A previous study that ran the same methodology, however, found significant correlations between these metrics and confusion (Andres et al., in press). The study reported confusion to be negatively correlated with earning a gold badge, but positively correlated with earning a silver badge.

3.4.2 Fine-Grained LTs and Affect

This sequence mining analysis will take into consideration the common sequences mined in the previous analysis (in 3.2), and correlate them with the percentages of time the students were observed to be either confused or bored. We hypothesize that some sequences will be characteristic of either affective state, and can then be used as indicators within the game. As with the analysis in 3.2, this analysis is still in progress.

4. Discussion, Conclusions, and Future Work

The study attempted to identify learning and affect trajectories among students using an educational game for Physics, called Newton's Playground. In each level of the game, players are made to get a green ball to a red balloon using one or a combination of these four simple machines: lever, ramp, springboard, and pendulum.

The study operationalized learning trajectories (LTs) on two levels: coarse-grained LTs track the students' performance in terms of gold and silver badges earned, and fine-grained LTs track in-game events that occur as students interact with the game. Four coarse-grained learning trajectories were analyzed, one for each of the four simple machines in the game. All four coarse-grained LTs showed a common pattern of students performing more poorly as time progressed, earning less badges, and solving less levels. The fine-grained LT analysis is still in progress.

The study also looked at boredom and confusion trajectories among the students. Results showed that while confusion was experienced by a steady amount of students throughout the 2 hours of gameplay, the percentage of students experiencing boredom increased over time.

The study attempted to find relationships between the learning trajectories and affect, and in doing so, found no significant relationships between performance and boredom. A previous study found significant correlations with confusion, however, where confusion was negatively correlated to earning a gold badge, and positively correlated with earning a silver badge (Andres et al., in press). The analysis between fine-grained LTs and affect is still in progress.

We speculate that there are a number of relationships that are worth further exploration. ICT has not successfully penetrated the education system in the Philippines. Several infrastructural, financial, and implementation hindrances still exist, and despite the government's best efforts to work around them, programs and projects still fall through the cracks. Several government projects are currently in place, however, that aim to 1) ease the integration of ICT in the classroom for both teachers and students, 2) help alleviate poverty, most of which harness technology to maximize outcomes, and 3) utilize technology to reach potential learners who don't have immediate access to any form of formal learning.

On the student level, poor prior knowledge (as evidenced by students' poor pre-test results) might have made the game daunting, causing the students' poor performance in the game over time. The game interaction time of two hours may have been too long, leading to the increase in boredom. Indeed, the researchers noticed that the students rushed through the post-test, implying that they wanted to leave the testing area as quickly as possible. Boredom might have led to systematic guessing and other similar non-learning behaviors, leading in turn to poor post-test scores (Baker et al, 2010). In future work, we intend to verify which among these hypotheses the data support. In doing so, we hope to contribute to principles that guide the development of good educational games.

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 for making this study possible. We would also like to thank Ryan Baker, Valerie Shute, Luc Paquette, Matthew Ventura, Matthew Small, and the Gates Foundation Grant #OP106038 for collaborating with us.

References

Andres, J.M.L., Rodrigo, M.M.T., Sugay, J.O., Baker, R.S., Paquette, L., Shute, V.J., Ventura, M., & Small, M. (in press). An Exploratory Analysis of Confusion Among Students Using Newton's Playground. 22nd International Conference on Computers in Education

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.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. Mathematical thinking and learning, 6(2), 81-89.

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. (2012). Dynamics of affective states during complex learning. Learning and Instruction, 22(2): 145-157.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). Learning Trajectories in Mathematics: A Foundation for Standards, Curriculum, Assessment, and Instruction. CPRE Research Report# RR-68. Consortium for Policy Research in Education.

Fisherl, C. D. (1993). Boredom at work: A neglected concept. Human Relations, 46(3), 395-417.

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.

Rodrigo, M. M. T., Baker, R. S., & Nabos, J. Q. (2010). The relationships between sequences of affective states and learner achievement. In Proceedings of the 18th International Conference on Computers in Education (pp. 56-60).

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.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. Journal for Research in Mathematics Education, 26, 114–145.

Wixon, M., d 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.