Chapter # - will be assigned by editors

SERIOUS GAMES ANALYTICS TO MEASURE IMPLICIT SCIENCE LEARNING

Elizabeth Rowe, EdGE at TERC

Jodi Asbell-Clarke, EdGE at TERC

Ryan S. Baker, Teachers College, Columbia University

Abstract: Evidence Centered Game Design (ECgD) is an increasingly popular model used for stealth game assessments employing education data mining techniques for the measurement of learning within serious (and other) games (GlassLab, 2014). There is a constant tension in ECgD between how pre-defined the learning outcomes and measures need to be, and how much important, but unanticipated, learning can be detected in gameplay. The EdGE research team is employing an emergent approach to developing a game-based assessment mechanic that starts empirically from what the players do in a well-crafted game and detects patterns that may be indicate implicit understanding of salient phenomena. Implicit knowledge is foundational to explicit knowledge (Polanyi, 1966) yet is largely ignored in education because of the difficulty measuring knowledge that a learner has not yet formalized. This chapter describes our approach to measuring implicit science learning in the game, *Impulse*, designed to foster an implicit understanding of Newtonian mechanics using a combination of video analysis, game log analyses, and comparisons with pre-post assessment results. This research demonstrates that it is possible to reliably detect strategies that demonstrate an implicit understanding of fundamental physics using data mining techniques on user-generated data.

Key words: Implicit Learning, Science Learning, Assessment, Educational Data Mining

1. INTRODUCTION

Games have long been recognized as natural assessments (Gee, 2003, 2007). However, it was the call for games as *stealth assessments* (Shute, Ventura, Bauer, Zapata-Rivera, 2009) that encouraged game-based learning researchers to think more about switching from using formal pre-post assessments to using assessments embedded within and/or consisting solely of gameplay data. In this move to stealth assessments, most instantiations use an Evidence-Centered Game Design (ECgD) model (Shute et al., 2009; GlassLab, 2014; Plass et al., 2014; Halverson, Wills & Owens, 2012) where explicit learning outcomes and measures are designed and developed as part of the game design process. The EdGE research team builds upon the ECgD framing with an emergent approach to detect implicit learning from complex patterns within data generated from a game whose mechanics are grounded in science. Grounded in videos of learners playing the game, EdGE studies where students' strategic game behavior is consistent with an implicit understanding of the science content and validates the use of those strategies against an external measure of implicit science learning (Asbell-Clarke & Rowe, 2014; Asbell-Clarke, Rowe & Sylvan, 2013). Implicit science learning is expressed in brief instances of play but unfolds and changes over course of play. This chapter outlines the theoretical lenses with which we view game-based science learning and describes the methods we use to measure that learning.

2. IMPLICIT SCIENCE LEARNING IN GAMES

Implicit knowledge (also called tacit knowledge) has a variety of forms or definitions. Polanyi (1966), a philosopher and scientist, argued that tacit knowledge is foundational to all explicit knowledge. Within tacit knowledge, Collins (2010) distinguishes between somatic tacit knowledge of primal tasks such as walking and talking; collective tacit knowledge in a community such as language and humor; and tacit relational knowledge, the tacit knowledge that with effort can become related to explicit, or formalized, knowledge. Tacit relational knowledge is likely of most direct consequence to formal education.

The ways in which implicit knowledge can impact learning and teaching is not completely new to education. Vygotsky (1978) described *preparedness for learning* as the abilities and understandings a learner brings to a learning situation that can be scaffolded by a teacher, environment, and tools. Late in the last century much literature in US science education turned attention to

implicit learning in the form of misconceptions that may get in the way of a learner's conceptual development (e.g. McCloskey, 1983a, 1983b; Minstrell, 1982). diSessa (1993) notes the robustness of physics misconceptions with over half of respondents agreeing with several common misconceptions about basic physics, such as Newton's Laws of Motion. diSessa also distinguishes between the intuitive knowledge that novices hold — that a book will not fall through a table or that a glowing filament is hot —from an expert understanding of these phenomena. For novices these understandings guide behavior, but are not necessarily expressible in formalisms or questioned in a deeper sense. Experts, however, not only think about a phenomenon in a more nuanced sense but also may seek consistency across phenomena to be able to abstract their experiences towards more general principles about the world (diSessa, 1993).

Implicit knowledge is, by definition, largely unexpressed by the learner making it particularly challenging to measure. Games may provide an innovative assessment solution as a growing body of research shows how games may engage learners in cognitive processes that are not necessarily perceived by learner or recognized in external learning assessments (Gee, 2013; NRC, 2011; Thomas & Brown, 2011; GlassLab, 2014).

The unique affordances that games offer for the measurement of implicit science learning include (a) the ability to engage learners by encouraging them to dwell in scientific phenomena over repeated trials towards success (with appropriate scaffolding and feedback) and (b) the wealth of information that can be recorded during game play to provide evidence of their implicit learning. These features open opportunities to reveal tacit learning previously invisible to educators.

3. STEALTH ASSESSMENTS

In the past decade, researchers have begun assessing learning occurring in interactive environments such as games (Shute & Ventura, 2013; Fisch et al., 2010; Halverson, Wills & Owens, 2012). A common way researchers have assessed learning in games is through pre-post tests or tasks before and after a specified period of gameplay. In contrast, *stealth assessments* measure learning using tasks embedded within the gameplay itself to "support learning, maintain flow, and remove (or seriously reduce) test anxiety, while not sacrificing validity and reliability" (Shute et al., 2010, p. 10). To satisfy validity and reliability requirements, researchers often use an Evidence-

Centered Design (ECD) framework that seeks to establish a logically coherent, evidence-based argument between the domain being assessed and assessment task design and interpretation (Mislevy & Haertel, 2006).

GlassLab (2014) describes how their team applied the ECD framework to the assessment of learning in SimCityEDU, creating an Evidence Centered Game Design (ECgD) approach that carefully defines how game and assessment design must work in concert to produce an evidentiary model for learning with an explicit framework for characterizing that evidence. Other researchers have developed stealth assessments guided by the ECgD framework using educational data mining techniques to discern evidence of learning from the vast amount of click data generated by online science games and virtual environments such as *Progenitor X* (Halverson, Wills & Owens, 2012), *EcoMUVE* (Baker & Clarke-Midura, 2013), *Newton's Playground* (Shute, Ventura & Kim, 2013), and *Surge* (Clark, Nelson, Chang, D'Angelo, Slack, & Martinez-Garza, 2011).

Within ECgD, measures of learning must be considered and designed along with the game mechanics. Plass and his colleagues (2014) argue that game mechanics, learning mechanics, and assessment mechanics must be designed in symbiosis with each other. For example, game mechanics to launch projectiles or maneuver objects through gravitational fields may have as their learning mechanics the development of specific understanding of forces and motion. The assessment mechanics in this case are the game behaviors (and often achievements) that correspond to the consistent use of strategies to grapple successfully with the forces and motion created by the gravitational effects.

However, there is a constant tension in ECgD to make sure gameplay is designed to support and measure meaningful learning, while also remaining open to important learning that may occur during gameplay but that designers may have not considered from the start. This is especially important in game spaces with hundreds or, in some cases, thousands of play patterns where the player can be successful. There can also be a tension between the most enjoyable game mechanics, and the most effective learning and assessment mechanics.

EdGE seeks to remain as open as possible to emergent evidence of implicit learning in games while still pursuing the logical coherence of the ECgD framework. We do this through a more open-ended, bottom-up iterative design process that optimizes game design for learner engagement (i.e. would they choose to play this game in their free time?) and allows the

assessment mechanisms to emerge from observations of gameplay rather than place any constraints on the game design. The remainder of this chapter describes the EdGE research team's attempt to push game assessment mechanic development towards that more emergent end of the spectrum while maintaining validity and reliability. We describe this process in the context of the game, *Impulse,* which has been played or downloaded by over 10,000 players online and through the iOS and Android app stores.

4. IMPULSE

EdGE designed *Impulse* to foster and measure implicit learning about Newton's First and Second Laws of motion by placing a simple game mechanic (get your particle to the goal without crashing into other particles). *Impulse* immerses players in an n-body simulation of gravitationally interacting particle in which they must predict the Newtonian motion of the particles to successfully avoid collisions and reach the goal (see Figure 1). For a better understanding of this work, readers are encouraged to play *Impulse* at edgeatterc.com/edge/games/impulse/.

Figure 1: A screenshot from *Impulse*. The player is the green particle and is going towards the cyan goal in the bottom-left corner.

The motions of all particles in the game obey Newton's laws of motion and gravitation, including accurate gravitational interactions and elastic collisions among ambient particles with varying mass. Players use an impulse (triggered by their click or touch) to apply a force to particles. If the player's particle collides with any ambient particle, the level is over and they must start again. Each level of the game gets more complex, requiring players to grapple with the increasing gravitational forces of an increasing number of particles and also particles of different mass (and thus inertia). For each level, they must accomplish this goal with 20 impulses. Each impulse depleted the energy available to the player in the game (measured by the green bar in the upper right corner of Figure 1). Once they exceed 20 impulses, the player no longer has energy left to apply any force to the particles.

Newton's First Law states that an object in constant motion will stay in constant motion unless acted upon by an external force. This is counterintuitive for many learners because we rarely encounter a frictionless environment in real life (McColskey, 1983). Newton's Second Law states that the acceleration an object experiences from a force depends on the mass of the object. The *game mechanic* increases n, the number of particles, to increase the complexity and difficulty of each level and also uses particles of different mass to provide opportunities for players to grapple with phenomena governed by Newton's First and Second Laws. The *learning mechanic* is designed assuming that as players dwell in increasingly complex situations in the game, they may build strategies that help build tacit knowledge that is foundational for explicit learning of the behaviors governed by these laws.

The *assessment mechanic* is designed to measure players' behaviors that may indicate they are gaining an implicit understanding of Newtonian motion. We look for patterns of play in the game data logs reflect behaviors that players demonstrate that are consistent with implicit understanding. For example, players may let a ball "float" with added force, and then use an opposing force to stop the balls motion—both consistent with an understanding of Newton's first law of motion. Even more directly, a player might consistently use more force to accelerate a heavier object than a lighter one – demonstrating an implicit understanding of Newton's second law.

5. ASSESSING IMPLICIT SCIENCE LEARNING

The EdGE research team is taking three steps to build assessment mechanics of Newton's First and Second Law for *Impulse*. First, we coded videos in terms of specific strategic moves, noting which strategic moves are consistent with an understanding of Newton's first and second laws (i.e., the phenomena in which they are dwelling). Second, we mined the game log data for evidence consistent with an implicit understanding of those laws. Finally, we will be validating those play patterns against learner performance on a pre-post assessment of those concepts. These steps vary slightly for each of Newton's Laws. While evidence for Newton's First Law can be found in a player's single actions (clicks), evidence for Newton's Second Law relies on the relationship between sequences of actions (i.e., how many times they click on particles of different masses within a short time).

We hypothesize that advancing to higher levels in *Impulse* depends upon, fosters, and demonstrates an implicit understanding of Newton's laws. While navigating among particles that are colliding and are attracted or repelled by each other, players need to "study" the particles' behavior. They must predict the motion of the particles so that they can avoid them as they travel to the goal. Specifically, we expect players to increase their understanding that each particle will keep moving on its path without an impulse or force from another particle (Newton's First Law) and that different mass particles react differently to the same force (Newton's Second Law).

5.1 Video Coding as Ground Truth

Two researchers, one the game designer with a physics background and the other with expertise in the learning sciences and limited background in physics, began developing the coding system using video recordings from two play test sessions, one with 10 high school students from urban and suburban schools in the northeastern U.S., and another with 6 Physics graduate students in from a small university in Canada. These samples represent players with novice and near-expert understandings of Newton's Laws of Motion.

Players' interactions with *Impulse* were recorded with Silverback software (ClearLeft, 2014) capturing both players' onscreen game activities and video of their faces and conversations. Students were asked to 'think aloud" while playing. Typically students played in groups, one student per computer,

prompting conversation about gameplay and phenomena they observed. Silverback solves many synchronization problems others have experienced using multiple video cameras to record screen activity, facial expressions, and conversations.

Data from a larger number of learners were needed to build detectors based on this coding system. These data were collected over six hour-long workshops conducted in March-June 2013 with 69 high school students (29 female) from urban and suburban schools in the Northeastern United States. A third coder with no physics background was trained using the coding system and coded randomly selected three-minute segments from all 69 videos. Segments were randomly chosen above Level 20 whenever possible to ensure players had already mastered the game mechanic and had encountered particles of different masses. Twenty-nine of the players (42%) did not reach level 20 and had time segments earlier in the game. Two additional coders and one of the designers of the coding system double coded the segments from 10 videos for inter-rater reliability checking.

The final version of this coding system presented here was developed through repeated coding of hundreds of clicks with different play styles. These codes are not mutually exclusive (i.e., it is possible for one click to be both a 'Float' and a 'Move Toward Goal'). Each click was coded with at least one of these codes. Table 1 includes definitions of the codes with inter-rater (human-human) Kappas exceeding 0.70 and the implicit understanding of Newton's First Law we claim they reflect.

Table 1. Video Codes, Definitions, and Kappas for Newton's First Law (NFL).

Intended Strategy Code label	**Game-based move**	**Implicit Understanding**	**Kappa**
Float	The learner did not act upon the player particle for more than 1 second	Player particle will move in a straight path if no force is applied (NFL)	0.759
Move Toward Goal	The learner intended to apply force to direct the player particle toward the goal	Control movement of player particle by applying force	0.809
Stop/slow down	The learner intended to use opposing force on player particle in the path of the player particle to stop/slow it down	Slow particle down by using an opposing force (NFL)	0.720

Keep player path clear	The learner intended to apply force to non-player particles to keep them out of the path of the player particle	Player particle will move in a straight path if no force is applied (NFL)	0.819
Keep goal clear	The learner intended to apply a force to non-player particles to keep the goal clear by removing the non-player particle	Control movement of non-player particles by applying force	0.832
Buffer	The learner intended to apply a force between the player and other particles to avoid collision	Control movement of player and non-player particles by applying force	0.772

Source: Rowe, Baker, Asbell-Clarke, Kasman, & Hawkins (2014).

When coding we distinguished between intended and actual game moves—what the player wanted to accomplish with each click versus what actually happened. Player intentions are judged based not only on their screen actions, but also audio commentary and mouse over behaviors. Often players hold their mouse over spots, ready to click if needed, providing visible clues of their intended path or strategy. While not directly visible in the clickstream data, these behaviors are observable in video and aid interpretation. For actual moves, we coded whether or not intended and actual moves matched and, if not, which of five unanticipated outcomes occurred. These unanticipated outcomes include (1) no effect on the target particle; (2) rapid acceleration of the target particle (i.e. click was too close to the particle and made it accelerate more rapidly than expected); (3) moved the player particle closer to another particle (i.e. causing a potential collision), (4) moved the player particle away from the goal (in the absence of reason to do so); and (5) the target particle did not move as expected with no negative consequences as is the case with the other outcomes. The reliability of this code depends on the reliability of the intended codes. If they did not agree on the intended strategy, it is likely they would not agree whether the actual move was as intended or not. Therefore, it was not surprising that the coding of unanticipated outcomes (Kappa=0.35) was much less reliable than the coding of intended moves (see Table 1).

Players clearing a particle from their path towards the goal may show evidence of their implicit understanding of Newton's First Law in that are predicting that the particle will stay at constant motion in the absence of a force (and thus will collide) so they impart the force to move it away. Even

more compelling evidence of an implicit understanding of Newton's First Law is when the player directly opposes straight-line motion with their impulse (Stop/Slow Down), explicitly providing the force needed to stop their particles' motion. When a player uses a Float strategy, particularly when accompanied by a mouseover trailing along with the particle, their behavior is consistent with an implicit understanding that an external force is not needed to keep the particle moving at a constant speed (Newton's First Law).

For evidence of an implicit understanding of Newton's Second Law, we coded information about the target of the click and whether or not the target of the current click was the same as the previous click (see Table 2). Together these codes were used to determine if the player treated the different mass balls differently, more specifically if they consistently used more force (clicks) to move the heavier particles than the lighter ones.

Table 2. Video codes, definitions, and Kappas used for measuring Newton's Second Law.

Code	**Definition**	**Kappa**
Target	Type of particle (player, other, both) the learner intended to move	0.920
Same as Last Target	The learner intended to move the same target as the last action	0.869

Source: Rowe, Baker, Asbell-Clarke, Kasman, & Hawkins (2014).

There were four different colored particles besides the player with each color signifying a different mass (in order from least to most massive): blue, red, white, dark grey. The color of the target was recorded alongside the target. The blue, red, and white balls also increased in size (consistent with the same density of ball) but the grey ball was most massive and smallest in size. This was to ensure that mass was being differentiated in players' behaviors rather than size. From these codes, the number of consecutive clicks for each color target was calculated.

5.2 Game Log Analyses

As the learner plays *Impulse*, the game logs every game event as well as the location of every object in the game space. Recorded game events include level starts/ends, pausing and resuming the game, clicks (impulses) in the game space, collisions between particles, collisions between the particles and the walls of the game space, and collisions of the player with the goal. The game state is recorded along with the event. The final outcome of each game

level is also recorded: Advance with energy remaining, Advance without energy remaining, Collision with energy remaining, Collision without energy remaining, Restart, and Quit. Players have a limited amount of energy (20 clicks) to at each level of the game, so if they 'Advance without energy remaining' it means they floated into the goal after they ran out of energy.

From this raw game log, we have distilled a set of 60+ features in five major categories: (1) Location/Vector Movement of Player Particle; (2) Timing and Location of Impulses; (3) Number and Location of Other Particles; (4) Overall Game Characteristics, and (5) Game Outcome. The feature distillation process explicitly selected features thought by domain experts to be semantically relevant to the strategies observed by the human coders (Sao Pedro et al., 2012). Table 3 gives a non-exhaustive list of examples:

Table 3: Distilled feature categories, examples, and rationale

Category	Distilled Feature Examples	Rationale
Player Particle		
1	Distance between Player and Goal	Players use different strategic moves when close to the goal than when farther away
2	Current speed of player particle	When the player is moving faster they need to use different strategic moves than when slow
3	Distance travelled since last event	This provides an indication of how much the game state has changed
4	Change in angle between player's path and a straight-line path to goal	Strategic moves vary depending on whether or not player has a straight-line clear path to the goal
Impulses		
1	Proximity of impulse to player particle	Identifies the likely intended target (player particle or other) of the impulse.
2	Time since last impulse	Very quick actions may indicate panicking or intentional increased force; very slow actions may indicate floating strategies

3	Distance from impulse to three closest other particles and their color	Identifies the likely intended target (player particle or other) of the impulse and identifies if players click more near certain color particles.
Other Particles		
1	Number of other particles in play space	Describes the potential complexity of the play space
2	Number of particles in path between player and goal	Describes difficulty of immediate task of getting to goal
3	Number of particles in current path of player particle	Describes immediate danger of collision
Overall Game Characteristics		
1	Total time spent playing this level across multiple rounds	Describes difficulty of the level
2	Total number of times playing this level	Describes players experience with the level

Source: Asbell-Clarke, Rowe, Sylvan, Baker (2013).

The distilled features were added to the original backend data. Using the synchronized timestamps, these features are then aggregated at the click level to map to the labels provided by the video coder (Sao Pedro et al., 2013).

5.2.1 Building Detectors of Strategic Moves: Evidence for Newton's First Law

With the distilled data and the human-coded data, we followed a standard process for developing a model that could replicate the human judgments using the distilled log files. In other words, the goal of these analyses was to develop software that could look at the logs of student interaction with the software, and come to the same judgments as a human being.

Specifically, we developed classifiers that could infer the human-coded data (1 for the presence of a specific category, 0 when it was absent), in

RapidMiner 5.3. A separate classifier was developed for each human-coded construct (strategic move), six classifiers in total.

Four algorithms were tried for the first three classifiers developed:

- W-J48—a "decision tree" algorithm which makes a set of yes/no decisions based on the data to make an eventual decision with a known confidence; based on the first decision, the second decision will be different (Quinlan, 1993)
- W-JRip—a "decision rules" algorithm which makes a set of yes/no decisions based on the data to make an eventual decision with a known confidence; the order of decisions is always the same regardless of previous decisions
- Logistic Regression—regression conducted using a logistic function in order to predict a binary variable rather than the quantitative variable predicted in linear regression
- Step Regression—regression conducted using a step function rather than a logistic function or a linear function using the standard software RapidMiner 5.3 with the Weka Extension Package. Step regression is not to be confused with stepwise regression.

These algorithms were selected based on their success in past problems where researchers attempted to classify student behavior within online learning environments for science inquiry (cf. Baker & Clarke-Midura, 2013; Sao Pedro et al., 2012, 2013; Baker et al., 2014), as well as in other domains. W-J48 worked best for the first three constructs, and so W-J48 was the only algorithm attempted for the remaining three. W-J48 is a decision tree algorithm with several virtues: it produces relatively interpretable models, is fast to create and use (facilitating both validation and use in a running system), and tends to be conservative (reducing the risk of over-fitting, where a model is fit to the noise in the data as well as the signal).

The models were validated in the following fashion. For each construct, the algorithm was validated using 4-fold student-level cross-validation. The students were randomly distributed into four groups. The algorithm was run, training a model on data from three of the groups. Then the model was applied to the data from the students in the fourth group, and tested to see how well the model functioned on this unseen group. It is important to use student-level cross-validation to avoid training and testing a model on the same student; if a student's behavior is idiosyncratic, then the model may become over-fit to that student and less able to function effectively for other students. Student-level cross-validation penalizes models that over-fit to the

specific student. Within student-level cross-validation, the number of folds may lie between 2 and the number of students. This type of cross-validation is thought to be asymptotically equivalent to the Bayesian Information Criterion (Moore, 2003); while the choice of number of folds remains arbitrary, four is a common number of folds that leads to models repeatedly being built on 75% of students and tested on the remaining 25%.

In this study, two goodness (performance) metrics were used to determine how effective each detector was: Cohen's Kappa (Cohen 1960) and A' (Hanley and MacNeil 1982). Each of these metrics was applied at the level of the three-minute segments coded from the video data.

Cohen's Kappa assesses the degree to which the detector is better than chance at identifying which segments involve a specific code. For example, a Kappa of 0.865 would indicate that a detector is 86.5% better than chance for a specific code. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly.

A' is the probability that the detector will correctly identify whether a specific code is present or absent in a specific clip, taking model confidence into account when comparing clips to each other. A' is equivalent to W, the Wilcoxon statistic, and closely approximates the area under the Receiver-Operating Curve (Hanley & MacNeil 1982). A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. For example, an A' of 0.967 indicates that a detector of "keep player path clear" can distinguish a student demonstrating that strategy within a 3-minute segment from a player not demonstrating that strategy, 96.7% of the time.

These two metrics have different virtues. Cohen's Kappa assesses the quality of a model's final decisions (and is therefore a better assessment of how well the model will perform when used to drive interventions in the most common fashion, assigning interventions when confidence is over 50%), while A' assesses a model's confidence in its decisions (and is therefore a better assessment of how well the model will perform when used in discovery with models analyses, which typically take percent confidence into account).

There are no specific cut-off values for the use of these metrics in educational data mining, as acceptable performance tends to depend on the usage and the expectations in the current domain; medical tests are published and used with A' values of 0.75-0.80 or higher; affect detectors are published as of this writing with Kappa values as low as 0.15 and A' values as low as

0.65 (Sabourin, Mott, & Lester, 2011; Pardos, Baker, San Pedro, Gowda, 2013). Kappa values above 0.5 and A' above 0.8 tend to represent state-of-the-art performance in most educational domains as of this writing.

Table 4: Kappas and A' for each Intended Strategic Move

Intended Strategic Move	**Kappa**	**A'**
Float	0.738	0.901
Move Toward Goal	0.757	0.907
Stop/Slow Down	0.512	0.779
Keep Player Path Clear	0.865	0.967
Keep Goal Clear	0.772	0.943
Buffer	0.759	0.928

Source: Rowe, Baker, Asbell-Clarke, Kasman & Hawkins (2014).

Table 4 shows the performance of the specific models created in this chapter. Hence, we have developed models that can judge a learner's strategic moves relevant to Newton's First Law, successfully drawing many of the same conclusions a human being can (for six codes). These models were assessed based on their ability to agree with a human rater on entirely new, unseen data, and achieve comparable reliability. They met this test, achieving reliability similar to the human coders (and much better than most automated detectors of this type in the published literature).

The ability to detect these strategic moves reliably in the game data logs means we can now compare the learning of those players who use these moves consistently to those who don't. We hypothesize that players who use these moves consistently will be better prepared to learn Newton's first law of motion in class having developed this implicit foundational knowledge.

5.2.2 Mining sequences of clicks: Evidence of Newton's Second Law

To seek evidence of implicit knowledge of Newton's Second Law of motion (F=ma), we analyzed sequences of fast clicks. In specific, we looked at the length of sequences where players clicked near each color particle to move

it. Each color of particle has different mass and size, represented by the different colors. By looking at how frequently the players click near the same particle in a short amount of time, we can see if they recognize that more massive particles require a greater degree of force to be moved the same distance – or if they confuse mass and size.

We examined this for a range of operationalizations of a "short time", e.g. fast clicking, treating the cut-off as being 1 second, 2 seconds, up to 10 seconds. The overall pattern of results was very similar across time lengths; within this chapter, we will just show values for 4 seconds, a time threshold long enough to include all students repeatedly clicking to move the same particle, but brief enough for students to avoid cases where the student is clicking on the same particle for different reasons. So, for each particle color, we looked for cases where a student clicked to move the same particle (as coded by the human coder) in under 4 seconds after the previous action. Then we look for how many times this happened in sequence (which would be 1 if the player clicked to move a particle once in under four seconds after the previous action and then did something else; 2 if the player clicked to move the same particle twice in under four seconds after the previous action and then did something else, and so on).

Within this analysis, we compared the sequence length for different particle colors, across all sequences. A between-subjects comparison was used, as different students played different levels and therefore received different particles (and some students did not click near all the particles they saw). This discards some within-subjects information leading to a conservative assumption (leading to less statistical power to find significant results). We compared each color particle to each other color particle, using a two-sample t-test. Then we applied the Benjamini and Hochberg (1995) post-hoc correction to control for having run six statistical tests. Benjamini and Hochberg is a "false discovery rate" post-hoc method that controls for the number of tests run while avoiding the over-conservatism that characterizes family-wise error rate methods such as the Bonferroni correction.

The Benjamini and Hochberg correction requires a smaller p value, for significance, varying by test (within this method, some tests in a set end up requiring a lower p value than others for significance). Three of the six differences between sequence length are statistically significant according to this test: grey versus red ($t(40)=5.25$, $p<0.001$, $\alpha = 0.008$), grey versus blue ($t(31)=3.76$, $p<0.001$, $\alpha = 0.017$), white versus red ($t(57)= 2.98$, $p=0.004$, $\alpha = 0.025$). A fourth was marginally significant, white versus blue ($t(48)= 2.07$, $p=0.04$, $\alpha = 0.03$). The remaining two tests were not significant, white versus

grey (t(37)=1.65, p=0.11, α = 0.042), and blue versus red (t(51)=0.49, p=0.63, α = 0.05). This pattern of results is more clearly shown in Figure 2.

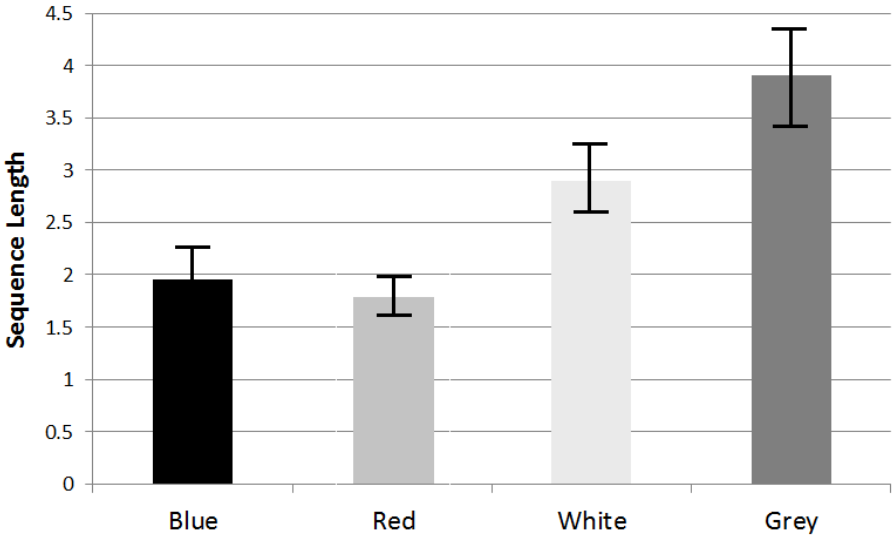

Figure 2. The average sequence length for the student quickly clicking each color particle. Standard error bars shown.

These findings show that players are markedly differentiating the particles in terms of their mass, which is consistent with an implicit understanding of Newton's second law of motion. In the game, the mass of the balls is near equal for the red and blue balls, and for the white and grey balls. Players behavior in the game are consistent with their differentiating these masses, they treat the red and blue ball similarly, but click more (impart more force) to accelerate the white and grey balls. Furthermore, the grey ball has a smaller radius of any of the other balls (as if it were made of a much more dense material) yet players still distinguish the mass from size as the factor causing the acceleration, demonstrating possible evidence of implicit understanding that the two particles have different relative density.

A second potential test of this is how far players click from the various particle colors, as closer clicks create a greater force on the object. We can compute this by looking at the distance the player was away from the particle when he or she clicked, with that particle as a target, and then computing a two-sample t-test with Benjamini and Hochberg adjustment (e.g. the same test as conducted immediately above) to compare between particles colors. In this case, we find that three of the six statistical tests are significant: red versus white, (t(89)=5.17, p<0.001, α = 0.008), grey versus

white ($t(49)=4.95$, $p<0.001$, $\alpha = 0.017$), and blue versus white ($t(33)= 4.82$, $p<0.001$, $\alpha = 0.025$). In other words, players always clicked further away from the white particle than the other particles. The remaining three tests were not significant: grey versus red ($t(68)= 1.36$, $p=0.18$, $\alpha = 0.03$), blue versus red ($t(86)=0.69$, $p=0.49$, $\alpha = 0.042$), and blue versus grey ($t(46)=0.65$, $p=0.52$, $\alpha = 0.05$). Therefore, there were no differences in click distance from the other particles. Note that the degrees of freedom are higher for these tests than for the previous set of tests; more students clicked near a particle of a certain color at least once, than clicked near that particle in under four seconds. The pattern of results for particle distance is more clearly shown in Figure 3.

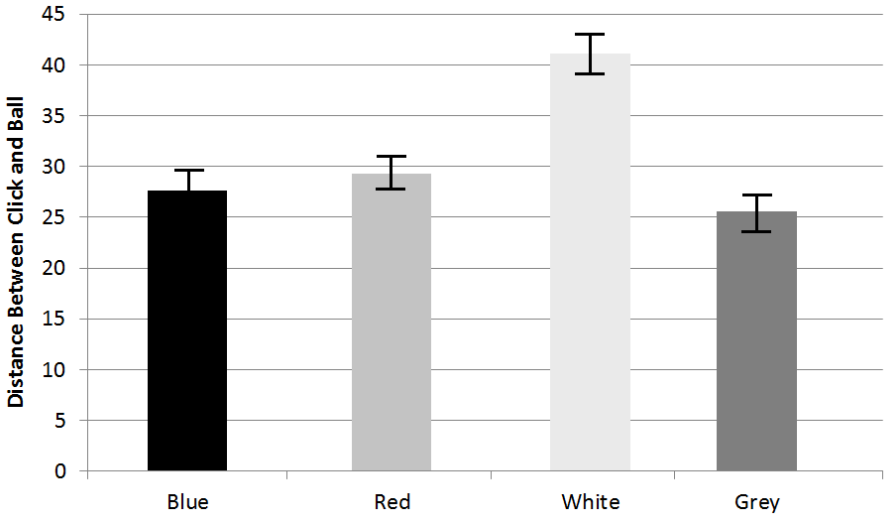

Figure 3. The average distance (pixels) away that the student clicked each color particle. Standard error bars shown.

Players treat most of the particles the same with regard to distance of the impulse, but the white particle appears to be an exception. This may likely be due to the larger radius of the white particle (it appears much larger than the other particles on the screen). This finding may be explained by the fact that the balls were in motion so players' accuracy in distance may have been compromised. The finding further highlights players' ability to distinguish that it is the mass of the ball, rather than the size, that is important in the relationship between force and acceleration.

6. DISCUSSION OF THIS APPROACH FOR SERIOUS GAME ANALYTICS

The results from this research provide a model set of methods to use game data logs to detect strategies that may be linked to foundational implicit knowledge that has previously gone unmeasured. We feel this emergent approach to developing a game-based assessment mechanic is particularly well suited to open-ended game spaces with large numbers of play patterns that could serve as evidence of implicit understanding. Table 5 provides a summary of how our method connects explicit learning outcomes to implicit game-based knowledge.

Table 5: Connecting explicit and implicit science knowledge

Explicit Learning Outcome	**Implicit Game-based Knowledge**	**Cognitive Strategy**	**Game-based Strategic Move**
Newton's First Law	Each particle will keep moving on its path without an impulse or force from another particle	Slow particle down by using an opposing force	Consistently click in the path of a particle, close enough to stop or slow it down
Newton's Second Law	The different mass particles react differently to the same force	Impart more force to move heavier particles than lighter particles	Consistently click more frequently next to heavier particles than lighter particles

We have shown that we can reliably detect a series of strategic moves in *Impulse* data that players were observed using in their quests to get their particle to the goal while grappling with Newtonian mechanics. The use of float, stop, and clear path strategies may indicate players' implicit understanding that the particle with stay in constant motion in the absence of an external force (Newton's First Law).

Even more striking to these authors is players' differentiation between masses of the particles in *Impulse.* The notable difference between clicks near light and heavy particles is a strong indicator of possible implicit understanding of Newton's Second Law. Players use more force to accelerate the heavier particles – even when they are smaller in diameter.

Having built and validated these detectors, we are now applying these detectors to a larger sample of gameplay data from 388 students as part of a

national implementation study of 39 classrooms (Rowe, Asbell-Clarke, Bardar, Kasman & MacEachern, 2014).

This user-generated data and distilled features will be inputted into RapidMiner, along with the previously generated W-J48 decision trees. The trees will be applied to the data, producing a prediction for every click of the probability that each of the relevant strategic moves in Table 3 were used. Every learner action in this game will be annotated with the probability that the learner was using each of the strategic moves.

We then plan to apply sequential pattern mining (Srikant & Agrawal, 1996) to the data set created by the application of the detector to all students' log data. The annotated logs will show us sequences of student strategic moves over time; sequential pattern mining will allow us to find out whether there are specific combinations of strategic moves that emerge over time and how those sequences are connected to broader learning of the physics concepts present in *Impulse*. Similar strategies have been used to infer whether students form strategies over time in Betty's Brain, a learning-by-teaching environment (Kinnebrew & Biswas, 2012).

Our ability to detect common strategies in the game data logs that are related to learning outcomes is a foundational step in research on implicit learning. Ultimately we are using these data along with many different instruments to measure engagement, attention, and non-cognitive factors that may be influencing the entire learning experience. In such, we are developing new models of learning in which data reveal learning that was previously invisible.

REFERENCES

Asbell-Clarke, J., Rowe, E., Sylvan, E., & Baker, R. (2013, June). Working through Impulse: Assessment of Emergent Learning in a Physics Game. Paper presented at the 9th annual meeting of the Games+Learning+Society (GLS) conference, Madison, WI.

Asbell-Clarke, J., Rowe, E., & Sylvan, E. (2013, April). Assessment Design for Emergent Game-Based Learning Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'13). Paris, France.

Asbell-Clarke, J & Rowe, E. (2014). Scientific Inquiry in Digital Games. In F. Blumberg (Ed.) *Learning by Playing: Video Games in Education*. New York: Oxford University Press.

Baker, R. S., & Clarke-Midura, J. (2013). Predicting successful inquiry learning in a virtual performance assessment for science. In *User Modeling, Adaptation, and Personalization* (pp. 203-214). Springer Berlin Heidelberg.

Baker, R. S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A., Metcalf, S.J. (2014) Extending Log-Based Affect Detection to a Multi-User Virtual Environment for Science. To appear in *Proceedings of the 22nd Conference on User Modelling, Adaptation, and Personalization*, 290-300.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Clark, D. B., Nelson, B., Chang, H., D'Angelo, C. M., Slack, K. & Martinez-Garza, M., (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers and Education,* 57(3), 2178-2195.

Clearleft Ltd. (2013) Silverback (Version 2.0) [Software]. Available from http://silverbackapp.com.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement 20 (1): 37–46. doi:10.1177/001316446002000104

Collins, H. (2010). *Tacit and explicit knowledge*: University of Chicago Press.

diSessa, Andrea A. (1993). Toward an Epistemology of Physics. Cognition and Instruction, 10(2/3), 105-225. doi: 10.2307/3233725

Fisch, S.M., Lesh, R., Motoki, E., Crespo, S., & Melfi, V. (2011). Children's mathematical reasoning in online games: Can data mining reveal strategic thinking? *Child Development Perspectives. 5*(2), 88-92.

Gee, J. P. (2003). What Video Games Have to Teach Us about Learning and Literacy. New York: Palgrave/Macmillan. 1st ed.

Gee, J. P. (2007). What Video Games Have to Teach Us about Learning and Literacy. New York: Palgrave/Macmillan. 2nd ed.

GlassLab (2014). Psychometric Considerations In Game-Based Assessment. Institute of Play. Downloaded 7/1/14 from: http://www.instituteofplay.org/work/projects/glasslab-research/

Halverson, R., Wills, N. & Owen, E (2012). CyberSTEM: Game-Based Learning Telemetry Model for Assessment. Presentation at 8th Annual GLS, Madison, WI, USA.

Hanley, J. A.; McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **143** (1): 29–36. PMID 7063747.

Hestenes, D., Wells, M., & Swackhamer, Gr. (1992). Force concept inventory. THE PHYSICS TEACHER, 30, 141.

Kinnebrew, J. S. and Biswas, G. (2012). Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. Proceedings of the International Conference on Educational Data Mining, 57-64.

McCloskey, M. (1983). Intuitive Physics. Scientific American, 248(4), 122-130.

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The physics teacher*, *20*(1), 10-14.

Mislevy, R. & Haertel, G. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Moore, A.W. (2003) Cross-validation for detecting and preventing overfitting. Statistical Data Mining Tutorials.

National Research Council (2011). *Learning Science Through Computer Games and Simulations.* M.A. Honey and M. L. Hilton (Eds.), Washington, DC: National Academies Press.

Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., & Gowda, S.M., (2013) Affective states and state tests: Investigating how affect throughout the school year predicts end of year

learning outcomes. Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, 117-124.

Plass, J., Homer, B.D., Kinzer, C.K., Chang, Y.K., Frye, J., Kaczetow, W., Isbister, K., Perlin, K. (2013). Metrics in Simulations and Games for Learning. In M. Seif El-Nasr, Drachen, A., & Canossa, A. (Eds.), *Game Analytics: Maximizing the Value of Player Data* (pp. 694-730). London: Springer-Verlag.

Polanyi, M. (1966). *The Tacit Dimension*. University of Chicago Press. Chicago, IL. USA.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann

Rowe, E., Asbell-Clarke, J., Bardar, E., Kasman, E., & MacEachern, B. (2014, June). Crossing the Bridge: Connecting Game-Based Implicit Science Learning to the Classroom. Paper presented at the 10th annual meeting of Games+Learning+Society in Madison, WI.

Rowe, E., Baker, R., & Asbell-Clarke, J., Kasman, E., & Hawkins, W. (2014, July). Building automated detectors of gameplay strategies to measure implicit science learning. Poster presented at the 7th annual Meeting of the International Educational Data Mining society, July 4-8, London.

Sabourin, J., Mott, B., and Lester, J. (2011). Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, pp. 286- 295.

Sao Pedro, M., Baker, R.S.J.d., Gobert, J. (2012) Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012), 249-260.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O. Nakama, A. (2013). Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction, 23* (1), 1-39.

Shute, V. J., Masduki, I., Donmez, O.... Wang, C-Y. (2010). Assessing key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, N. M. Seel (Eds.), Computer-based diagnostics and systematic analysis of knowledge (281-309). New York, NY: Springer-Verlag.

Shute, V. & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

Shute, V., Ventura, M. & Kim, J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. *The Journal of Educational Research*, *106 (6.*), 423-430, **doi:10.1080/00220671.2013.832970**

Shute, V., Ventura, M., Bauer, M., and Zapata-Rivera, D., (2009). Melding the power of serious games and embedded assessment to monitor and foster learning? Flow and Grow. *Serious Games: Mechanisms and Effects, 1* (1), 1-33.

Srikant, R., & Agrawal, R. (1996). *Mining sequential patterns: Generalizations and performance improvements* (pp. 1-17). Springer Berlin Heidelberg.

Thomas, D. and Brown, J. S. (2011). *A New Culture of Learning: Cultivating the Imagination for a World of Constant Change*. Lexington, KY: CreateSpace.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes Cambridge, Mass.: Harvard University Press.

ACKNOWLEDGEMENTS

We are grateful for NSF/EHR/DRK12 grant #1119144 and our research group, EdGE at TERC, which includes Erin Bardar, Teon Edwards, Jamie Larsen, Barbara MacEachern, Emily Kasman, and Katie McGrath. Our evaluators, the New Knowledge Organization, assisted with establishing the reliability of the coding.

AUTHOR INFORMATION

Elizabeth Rowe
EdGE at TERC
2067 Massachusetts Avenue
Cambridge MA 02140
Phone: 617-873-9704
Email address: elizabeth_rowe@terc.edu
Website: edge.terc.edu

Dr. Elizabeth Rowe is the Director of Research for the Educational Gaming Environments (EdGE) group at TERC, responsible for data collection, analysis and interpretation for all EdGE projects. In her 14 years at TERC, Dr. Rowe has studied and developed innovative uses of technology in and out of school including several NSF-funded projects such as *Kids' Survey Network, InspireData* software for K-12 students, and the *Learning Science Online* study of 40 online science courses for teachers. Dr. Rowe has led formative and summative evaluations of several technology professional development programs. Prior to joining TERC, Dr. Rowe was a research analyst at the American Institutes for Research where she analyzed national survey data for the National Center for Education Statistics. She holds a bachelor's degree in mathematics and a Ph.D. in human development and family studies.

Jodi Asbell-Clarke
EdGE at TERC
2067 Massachusetts Avenue
Cambridge MA 02140
Phone: 1-617-873-9716
Email address: jodi_asbell-clarke@terc.edu
Website: http://edge.terc.edu

Dr. Jodi Asbell-Clarke is the director of the Educational Gaming Environments Group (EdGE) at TERC in Cambridge, MA, USA. TERC is a

not-for-profit research and development organization that has been focusing on innovative, technology-based math and science education for nearly 50 years. As the director of EdGE, Jodi leads a team of game designers, educators, and researchers who are designing and studying social digital games as learning environments that span home, school, and community. Jodi's background includes MA in Math, an MSc in Astrophysics and a PhD in Education. She started her career at IBM working on the first 25 missions of the space shuttle as an onboard software verification analyst. After teaching at the laboratory school at University of Illinois, she joined TERC and has spent the past 20+ years developing science education programs and researching new ways to promote science learning. In 2009, she co-founded EdGE at TERC.

Ryan S. Baker
Teachers College
Columbia University
525 W. 120th St.
New York NY 10027 Box 118
Phone: (212) 678-8329
Email: baker2@exchange.tc.columbia.edu
Website: http://www.columbia.edu/~rsb2162/

Ryan Baker is Associate Professor of Cognitive Studies at Teachers College, Columbia University. He earned his Ph.D. in Human-Computer Interaction from Carnegie Mellon University. Dr. Baker served as the first technical director of the Pittsburgh Science of Learning Center DataShop, the largest public repository for data on the interaction between learners and educational software. He is currently serving as the founding president of the International Educational Data Mining Society, and as associate editor of the Journal of Educational Data Mining. His research combines educational data mining and quantitative field observation methods to better understand how students respond to educational software, and how these responses impact their learning. He studies these issues within intelligent tutors, simulations, multi-user virtual environments, and educational games.