Assessing Implicit Science Learning in Digital Games

Elizabeth Rowe[1] (corresponding author)
elizabeth_rowe@terc.edu

Jodi Asbell-Clarke[1]
jodi_asbell-clarke@terc.edu

Ryan S. Baker[2]
ryanshaunbaker@gmail.com

Michael Eagle[3]
maikuusa@gmail.com

Andrew G. Hicks[3]
Drew.G.Hicks@gmail.com

Tiffany M. Barnes[3]
tmbarnes@ncsu.edu

Rebecca A. Brown[3]
rabrown7@ncsu.edu

Teon Edwards[1]
teon_edwards@terc.edu

[1] Educational Gaming Environments (EdGE) at TERC

[2] Teachers College, Columbia University

[3] North Carolina State University

Corresponding author:
Elizabeth Rowe
Educational Gaming Environments (EdGE) group, TERC, Cambridge, MA 02140.
Email: elizabeth_rowe@terc.edu

Abstract

Building on the promise shown in game-based learning research, this paper explores methods for Game-Based Learning Assessments (GBLA) using a variety of educational data mining techniques (EDM). GBLA research examines patterns of behaviors evident in game data logs for the measurement of implicit learning—the development of unarticulated knowledge that is not yet expressible on a test or formal assessment. This paper reports on the study of two digital games showing how the combination of human coding with EDM has enabled researchers to measure implicit learning of Physics. In the game *Impulse,* researchers combined human coding of video with educational data mining to create a set of automated detectors of students' implicit understanding of Newtonian mechanics. For *Quantum Spectre*, an optics puzzle game, human coding of Interaction Networks was used to identify common student errors. Findings show that several of our measures of student implicit learning within these games were significantly correlated with improvements in external post-assessments. Methods and detailed findings were different for each type of game. These results suggest GBLA shows promise for future work such as adaptive games and in-class, data-driven formative assessments, but design of the assessment mechanics must be carefully crafted for each game.

Keywords: computer-based assessment, implicit science learning, game-based learning, educational data mining, learning analytics.

Assessing Implicit Science Learning in Digital Games

The assessment of student learning often relies on evidence-based measurement of salient knowledge and skills (NRC, 2011). Formalisms used in common learning assessments, such as text and equations, may present barriers to the expression of cognitive abilities demonstrated in many everyday activities (Sternberg, 1996). Many students may host a wealth of untapped knowledge that could be demonstrated and leveraged with appropriate tools such as digital games (Gee, 2007; Haladyna & Downing, 2004; Thomas & Brown, 2011). Data-driven game-based learning assessments show promise to measure implicit knowledge – knowledge that is not articulated and thus hard to measure with traditional assessments. This provides an innovative form of formative, computer-based assessments that teachers can leverage for improved classroom learning.

This paper introduces a model of Game-Based Learning Assessments (GBLA) to measure implicit learning. We introduce the theoretical lenses with which we view GBLA and implicit learning and then describe the study of two different physics learning games, *Impulse* and *Quantum Spectre*. They both use popular game mechanics embedded in authentic science simulation, but *Impulse* is a fast action game dealing with idealized physics that is different than everyday experiences, while *Quantum Spectre* is a slower-paced puzzle game that closely mimics everyday experience. We provide an overview of each game with the associated data mining methods used to measure implicit physics learning in the game, and then present the methods and findings from classroom studies that examined how students' implicit learning in the game related to scores on pre/post assessments of similar content. We conclude with a discussion of the implications of this research on design, teaching, and future game-based learning research.

Literature Review

Game-based learning research has captured the attention of many researchers because of the natural potential for assessment, or GBLA. Digital games engage learners in complex activity and researchers are able to track their behaviors and actions within that activity to inform the assessment of learning (Gee, 2007; Shute, Ventura, & Kim, 2013). Games compel players to persist in complex gameplay fostering high-level reasoning, inquiry, persistence, and creativity (Asbell-Clarke, et al., 2012; Shute, Ventura, & Ke, 2015; Steinkuehler & Duncan, 2008; Qian & Clark, 2016). Games also promote STEM learning (Clark et al., 2011; Rowe, Asbell-Clarke, Bardar, & Edwards, 2017; Shute, Ventura, & Kim, 2013; Squire, 2011; Hamari, Shernoff, Rowe, Coller, Asbell-Clarke, & Edwards, 2015). GBLA has been framed as *stealth assessments* (Shute, Ventura, Bauer, & Zapata-Rivera, 2009) that aim to move away from external pre/post testing and towards assessments embedded within and/or consisting solely of gameplay data. GBLA often uses an Evidence-Centered Game Design (ECD) model (Shute et al., 2009; GlassLab, 2014; Plass et al., 2014; Halverson, Wills, & Owens, 2012). To build GBLA for implicit science learning, we draw from ECD but use educational data mining as the basis for inferring implicit physics knowledge not yet expressed formally by the learner.

Implicit Science Knowledge

The importance of implicit knowledge to science education is highlighted by the “misconception” literature in Physics education that focused on underlying misunderstandings that may interfere with conceptual development (diSessa, 1993; McCloskey, 1983a, 1983b; Minstrell, 1982). For example, many undergraduate physics students believed that objects will come to rest in the absence of a force (thus violating Newton’s first law of motion), or that an object at the end of a string being swung in a circle will continue to move in a circle even after the string is cut (e.g., McCloskey, 1983a, 1983b; Minstrell, 1982). When considering friction and gravity, learners may appropriately build these "misunderstandings" through evidence they gather while dwelling in their daily experiences since most experiences include frictional and gravitation forces they may not be aware of. Their mental models may be "correct" but are built upon a different body of evidence.

In their description of a Preparation for Future Learning (PFL) model that emphasizes the role of prior knowledge in the learning process, Bransford & Schwarz (1999) argue that assessments should measure how people know, not what they know. Similarly, Cook and Brown (1999) suggest differentiating between what learners can express from what learners are able to do, noting that studying learners' behaviors may provide a key lens to their implicit knowledge. Prior knowledge may be implicit—not yet articulated by the learner—which makes it difficult to assess (Reber, 1993; Collins, 2010). Implicit knowledge is thought to be foundational to learning (Polanyi, 1966; Lucariello, 2016; Brown, Roediger, & McDaniel, 2014; Underwood, 1996; Kahneman, 2011) and ignoring it may put many students at a disadvantage, particularly neurodiverse learners because explicit assessment may not capture what they are capable of doing and understanding (Haladyna & Downing, 2004). Bransford and Schwarz (1999).

GBLA: Stealth Assessment

Researchers are examining how stealth assessments within games may provide new quantifiable and scalable ways to measure learning, typically by studying patterns of behaviors that may provide evidence of game-based learning (de Klerk, Veldkamp, & Eggen, 2014; Kim, Almond, & Shute, 2016; Owen, Ramirez, Salmon, & Halverson, 2013; Fu, Zapata, &

Mavronikolas, 2014; Martin, Petrick, Forsgren, Aghababyan, Janisiewicz, & Baker, 2015; Riconscente, Mislevy, & Corrigan, 2015).

Most stealth assessments in GBL use an Evidence-Centered Design model (Shute et al., 2009; GlassLab, 2014; Plass et al., 2014; Halverson, Wills & Owens, 2012) to establish a logically coherent, evidence-based argument between the domain being assessed and assessment task design and interpretation (Mislevy & Haertel, 2006). Stealth assessments measure learning using digital 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).

However, ECD can be limited in cases where we are not entirely sure which behaviors indicate implicit knowledge. When humans can recognize behavior as demonstrating implicit knowledge, but it is impractical to identify exactly how they recognized the demonstration – for example, in cases where there are fuzzy boundaries between the behavior associated with having implicit knowledge and the behavior associated with not having implicit knowledge, or whether several factors can contribute to an assessment, ECD and other knowledge engineering methods can perform relatively poorly. In these cases, better results can sometimes be obtained through methods that leverage educational data mining to automatically determine the behaviors and cut-offs that are associated with human labels of a construct. For example, this approach has led to effective models that can recognize more complex (but still appropriate) science inquiry strategies than simple vary-one-thing-at-a-time strategies (Sao Pedro et al., 2013) and models that can infer whether a student using a virtual learning environment is conducting inquiry that places them on track to obtain the correct final conclusion (Baker & Clarke-Midura, 2013).

As such, we strive to build an evidence model drawing from ECD and grounded in behaviors that players utilize to grapple with scientific challenges they face in science learning games. However, because player paths towards solutions in *Impulse* and *Quantum Spectre* are open to many possible trajectories, we opted for research methods that could remain open to a very large number of possibilities of what players might do, using the more emergent method of data mining to help us model whether or not a player's behavior reflects appropriate implicit knowledge. As such, this work attempts to make a contribution to computer-based assessment, facilitating the assessment of implicit knowledge in complex learning games.

Research Design

This paper reports on the final validation step towards GBLA research in two implicit physics learning games. In each of these physics learning games, *Impulse* and *Quantum Spectre*, it is possible to view student's gameplay as presenting evidence of their implicit understanding of fundamental scientific laws. The purpose of this study of two physics learning games is to demonstrate that the gameplay behaviors we have identified as potential evidence of implicit science learning are, in fact, valid measures. For each game we examined the research question:

- How do players' patterns of behaviors exhibited in the game related to their changes in pre/post test scores on related content?

This work lays the foundation for using these automated measures to further scaffold student science learning through adaptive versions of the games, dashboards, and other tools designed to share student learning progress with teachers. To experience the games described below, readers are encouraged to play *Impulse* and *Quantum Spectre* at games.terc.edu/Impulse/ and games.terc.edu/QuantumSpectre/.

Game1: *Impulse*

Game & Learning Mechanic

Impulse was designed to foster and measure implicit learning about Newton's first and second laws of motion. *Impulse* immerses players within a simulation of an ideal form Newton's laws of motion without the distractions of a background gravitational field or friction. The behavior of the particles in the game (and thus the players' reactions) are consistent with the Newton's laws of motion as they are typically taught (and tested) in high school physics class, rather than the "misconceived" version complicated by friction and Earth's gravity that learners experience in their everyday life.

The game mechanic requires players to get their particle to the goal without crashing into other particles within a simulation of gravitationally interacting particles (see Figure 1). All the particles obey Newton's laws of motion. Players use an impulse (triggered by a click or touch) to apply a force to particles. Each level gets more complex, requiring players to grapple with the increasing gravitational forces of an increasing number of particles and also particles of different mass. The game was designed with the hypothesis that patterns of play in the game may show evidence of implicit learning of Newton's first and second laws of motion:

> **NFL**: An object will remain in constant motion (or rest) in the absence of a net external force.
>
> **NSL**: The acceleration of an object caused by a net external force is depending on its mass (F=ma). In layperson's terms, this means the heavier an object, the harder it is to move.

The *Impulse* particles have different mass, as distinguished by their color, including very dense particles that have a small radius but the heaviest mass. The text between levels of the game briefly introduces the particles as heavy or light but does not introduce any discussion of the related physics.

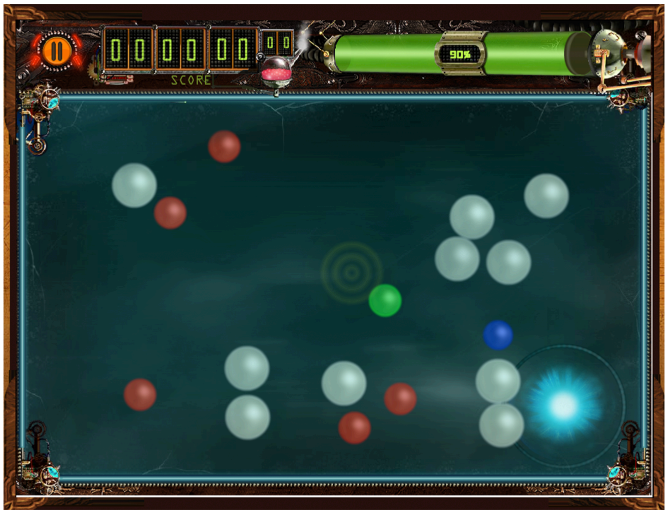


Figure 1: A screenshot from *Impulse*. The player is the green particle and is going towards the cyan goal in the bottom-right corner.

Building Implicit Game-Based Learning Assessments in Impulse

Three steps were taken to build GBLA of students' implicit understanding of Newton's First and Second Laws for *Impulse*. First, we coded videos in terms of specific strategic moves, noting which moves are consistent with an understanding of Newton's First and Second Laws. Second, we built detectors (i.e., classification algorithms) of the strategic moves consistent with an implicit understanding of those laws. Finally, in the results reported below, we examine the relationship between those strategic moves and learner performance on a pre/post assessment of those concepts to establish the validity of the in-game measures as assessments of implicit science learning.

Step 1: Human Coding. Using video and screen capture of gameplay from 69 high school students (29 female) from urban and suburban schools in the northeastern United States, we developed a coding system of strategic gameplay exhibited while playing *Impulse* during group think-aloud sessions (Rowe, Baker, Asbell-Clarke, Kasman, & Hawkins, 2014 and Rowe, Asbell-Clarke, & Baker, 2015). From the video analysis, several moves and strategies were evident in gameplay and described during playtesting sessions as potentially consistent with the salient phenomena of Newton's First and Second Laws of Motion. For example, players were observed to lift their hands from the keyboard and say "let it float" when allowing the particle to remain in constant motion without exerting a force. This understanding may also be consistent with their exerting an opposing (or buffering) force in the path of an incoming dangerous particle.

Table 1 includes definitions of the codes hypothesized to reflect an implicit understanding of Newton's First Law (NFL) with inter-rater (human-human) Kappas exceeding 0.70. A more detailed coding manual is available from the authors.

Table 1. Strategic Move Codes, Indicators, Implicit Science Understanding, and Kappas for Newton's First Law (NFL).

Intended Strategy Move Code	**Game-based move indicators**	**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 the 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)

As evidence of implicit understanding of Newton's Second Law, we looked for patterns in how players treated the different colored (and thus mass) particles. *Impulse* included 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 blue, red, and white particles also increased in size (consistent with constant density) but the grey particle was most massive and smallest in size. This was to ensure that mass rather than size was being differentiated in players' behaviors.

In the video coding we observed players commenting "Those white ones are hard to move" or "those blue ones really fly", supporting our hypothesis to analyze players' differentiation in how they treated the different mass particles, more specifically if they consistently used more force (clicks) to move the heavier particles than the lighter ones, which would be consistent with an implicit understanding of Newton's second law.

Step 2: Data Mining Detectors: Because *Impulse* logs every game event as well as the location of every object in the game space and feeds that data into the game data collection architecture, *Data Arcade*, we distilled the raw data to a set of 60+ features in five major categories: (a) Location/Vector Movement of Player Particle; (b) Timing and Location of Impulses; (c)

Number and Location of Other Particles; (d) Overall Game Characteristics; and (e) 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 (cf. Sao Pedro, Baker, & Gobert, 2012).

To examine players' behaviors related to Newton's first law of motion**,** we followed a standard process for developing a model that could replicate the human judgments using the distilled log files (cf. Baker & Clarke-Midura, 2013; Sao Pedro et al., 2012, 2013; Baker et al., 2014). The goal of these analyses was to develop detectors that can judge a learner's strategic moves relevant to Newton's laws of motion, successfully drawing many of the same conclusions a human being can. These models were assessed based on their ability to agree with a human rater on entirely new, unseen data, and achieve comparable reliability to human raters (see Rowe, Asbell-Clarke & Baker, 2015 for details regarding the reliability of these detectors). They met this test, achieving levels of reliability comparable with human raters, levels of reliability also comparable to medical diagnostics used to make real-world decisions about medical treatment (cf. Revell et al. 2013) as well as the diagnostics used in effective online learning platforms using detection of this nature (Sao Pedro et al., 2013).

Using the detectors to compare the learning of those players who use these moves consistently to those who do not, we hypothesize that we should be able to predict implicit learning of Newton's first law of motion. If the detectors are a valid form of GBLA, we hypothesized that those players who exhibit consistent use of Float, Stop/Slow Down, and Keeping the Player Path Clear strategies will have larger gains on an external pre-post assessment of Newtonian mechanics than players who use them less frequently.

To seek evidence of implicit understanding of Newton's Second Law of Motion (F=ma), we analyzed sequences of fast clicks by students. Aggregating data across all 69 students, we found students treated these pairs of particles differently: grey versus red, grey versus blue, white versus red, and white versus blue. The remaining two tests were not significant, white versus grey and blue versus red (see Rowe, Asbell-Clarke, & Baker, 2015 for more details). This pattern of results is more clearly shown in Figure 2.

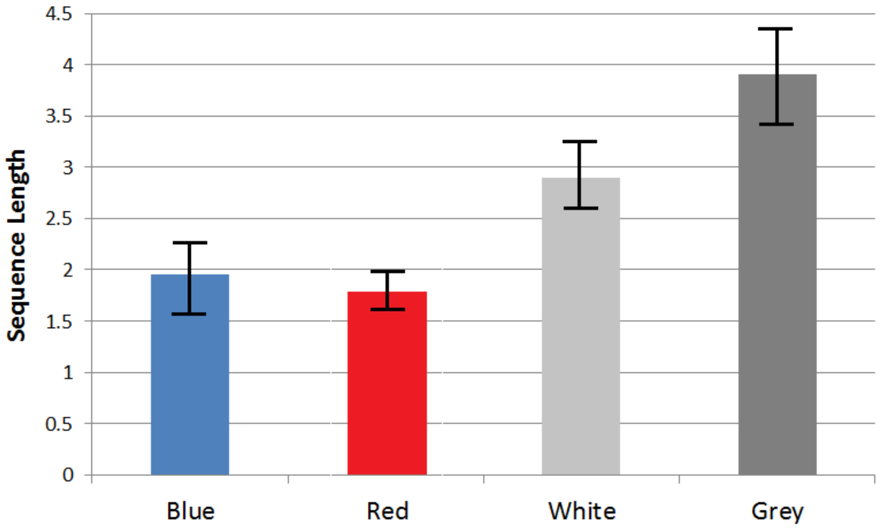


Figure 2: Sequence length (average number of clicks) by particle color, from least mass (blue) to greatest mass (grey). Source: Rowe, Asbell-Clarke, & Baker, (2015)

Across colors, this pattern suggests that the entire group of players was differentiating overall between particles with different masses, and thus this behavior of mass-differentiation could be used as a measure of implicit physics learning for individual players. The next step would be to validate it by comparing the prevalence of this behavior with players' scores on pre/post assessments.

Game 2: *Quantum Spectre*

Game & Learning Mechanic

When building GBLA *Quantum Spectre* is a puzzle-style game that uses scientifically accurate simulations of optical devices and colored laser beams to targets (see Figure 3). For each level, the player must place and rotate the lenses, flat and curved mirrors, and beam-splitters provided in the inventory to direct the laser to the matching colored target.

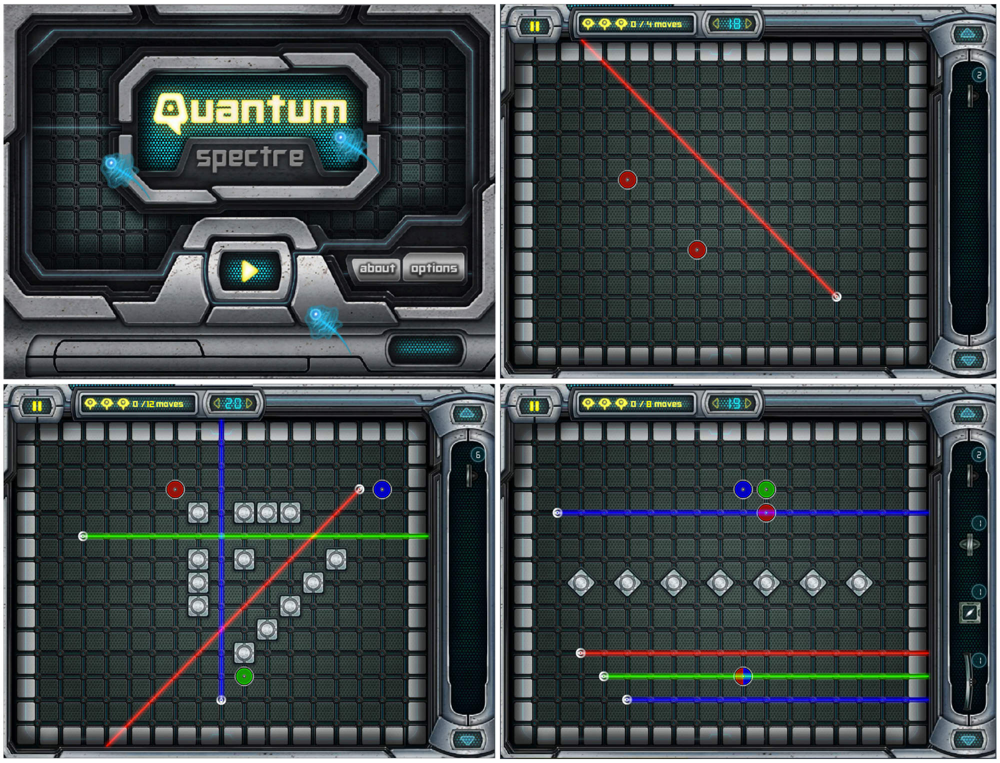


Figure 3: Quantum Spectre screenshots. Using mirrors and other optical devices from the right-side inventory, players must direct the colored laser beams to the matching targets.

Source: Hamari, Shernoff, Rowe, Coller, Asbell-Clarke, & Edwards (2015).

To examine implicit learning in *Quantum Spectre* we focus on two key science concepts:

- The **Law of Reflection**, which states that the angle of incidence of a light beam reflecting off a smooth surface equals the angle of reflection. To be consistent with understanding, players must correctly rotate the optical devices to solve the puzzles.
- **Slope**: Slope is the ratio of x and y coordinates when describing a line (rise/run). To be consistent with understanding, players must correctly place the optical devices on the puzzle grid to solve the puzzle.

Building Implicit Game-Based Learning Assessments in *Quantum Spectre*

Instead of extensive human coding, we sought a method that could reduce the thousands of different solutions sets players used for the puzzles to a manageable set of patterns that could be interpreted by humans. Thus, we focused on student step-by-step actions called *interactions*, and we aggregated all of the student gameplay interaction logs into an *Interaction Network* (IN) for each puzzle (Eagle Hicks, Peddycord, & Barnes, 2015). INs use complex network representation to represent student solution attempts. Game states are represented as vertices and game-actions are represented as edges with additional information preserved in the network elements. Once the networks were formed, we were able to apply data mining and visualization techniques in order to derive insight into the player behavior. This process is summarized below and can be seen in more detail in Rowe, Asbell-Clarke, Eagle, Hicks, Barnes, Brown, and Edwards (2016).

STEP1: Build Interaction Networks. To construct an Interaction Network, we collected the set of all solution attempts for each puzzle. Each interaction is defined as Initial State, Action, and Resulting State, from the start of the puzzle until the player solves the puzzle or exits the system.

A sample trace is shown in Figure 4.

Step 0 is the initial puzzle.

In Step 1 the player PLACES a mirror at (4, 1).

In Step 2 the player ROTATES the mirror 90 degrees.

In Step 3, the player PLACES another mirror at (2, 3).

In Step 4, the player ROTATES the mirror 270 degrees and solves the puzzle.

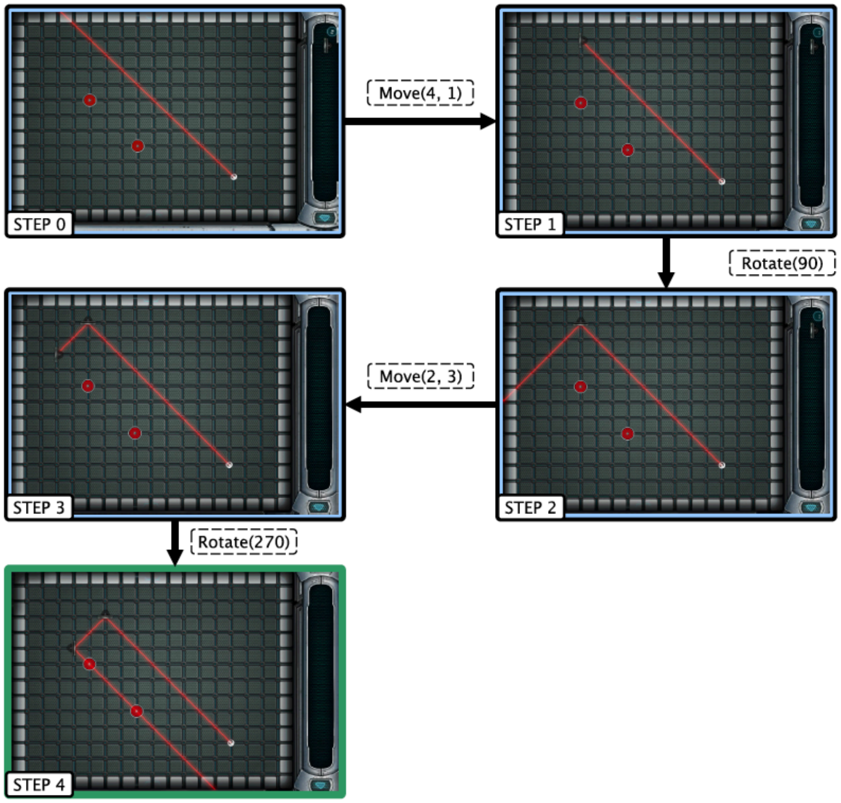


Figure 4: Sample trace of player actions in *Quantum Spectre* Puzzle 18. Each square represents an individual game state, and the edges represent actions that move a player from one state to another. Source: Rowe, Asbell-Clarke, Eagle, Hicks, Barnes, Brown, & Edwards (2016)

STEP 2: Clustering Paths by Shape. To simplify the state space, we consolidated game states where mirror and lens configurations resulted in similar output (the overall shape of the laser is functionally equal in terms of activating the sensors, etc.). Considering the output of a state as well as the position/orientation of objects in that state (Eagle & Barnes, 2014), we grouped equivalent states using *laser shapes* as a state representation. Figure 5 shows an example of a laser shape and its equivalent game states. While the shape of the laser's path (laser shape) is the same for each of these states, the difference is in which exact vertex the player placed the mirror.

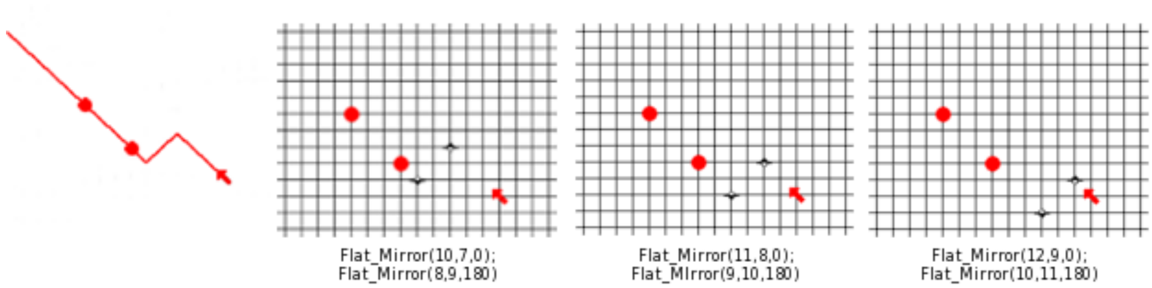


Figure 5: Using laser shape to group similar game states in Puzzle 18.

To further reduce the number of states and explore higher-level strategies of student solution attempts, we used a network clustering method called an Approach Map (Eagle & Barnes, 2014). The Approach Map identifies important actions (edges) that distinguish one group of IN regions from another (Eagle & Barnes, 2014). These groups were then reviewed by researchers and designers to align IN groups with those consistent with relevant game or science knowledge. This method enabled great simplification of the data analysis. For example, Puzzle 18 pictured above, for example, had 3458 unique edges (player actions) and 1800 unique game states, which were reduced to 22 groups of INs that could be human coded for further analysis.

STEP 3: Coding clusters of laser shapes. The game designer and researchers classified the laser shapes into one of four categories:

1) ***Correct move***—placement and rotation of the mirror are consistent with an eventual goal state.

2) ***Placement errors***—placement of the mirror in a location that does not match a goal state—may indicate a lack of understanding of slope.

3) ***Rotation errors***—rotation of a mirror to an angle that does not match a goal state—may indicate a lack of understanding of the Law of Reflection.

4) ***Puzzle errors***—placements of mirrors that were not consistent with a goal state, but were indicative of misunderstanding the game's puzzle elements (rather than the science concepts) (Eagle, Rowe, Hicks, Brown, Barnes, Asbell-Clarke, & Edwards, 2015).

The game designer and a researcher independently determined which placements for Puzzles 14-23 would constitute puzzle errors and completely agreed on the 33 puzzle errors. Puzzles 16, 21, and 23 had the largest number of potential puzzle error placements with 5, 7, and 8, respectively.

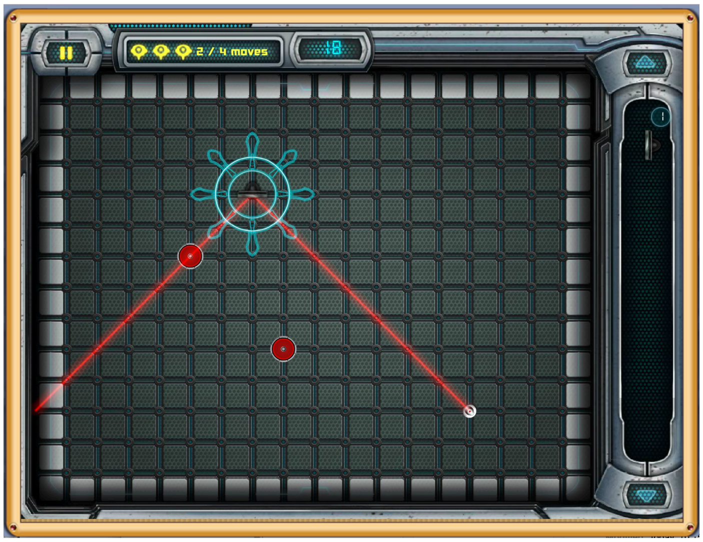


Figure 6: Sample Puzzle Error in Level 18.

For example, in the puzzle shown in Figure 4, a correct solution requires players to use the two available mirrors to direct the laser through the two targets simultaneously. In Figure 6 above, player actions are consistent with someone who understands slope (i.e., they placed the mirror on the path of the laser) and the Law of Reflection (i.e., they rotated the mirror to reflect the mirror through the target). However, these actions will not enable the student to solve this puzzle with one remaining flat mirror.

STEP4: Automated coding of errors. Once all laser shapes had been coded and puzzle error placements identified, we automated the coding of individual player actions. Every player action was classified as a Placement Error, Rotation Error, or Puzzle Error (0=Not Present; 1=Present). These are mutually exclusive player actions. Player actions with none of these errors were classified as Correct. Figure 7 shows the distribution of player behaviors across each puzzle.

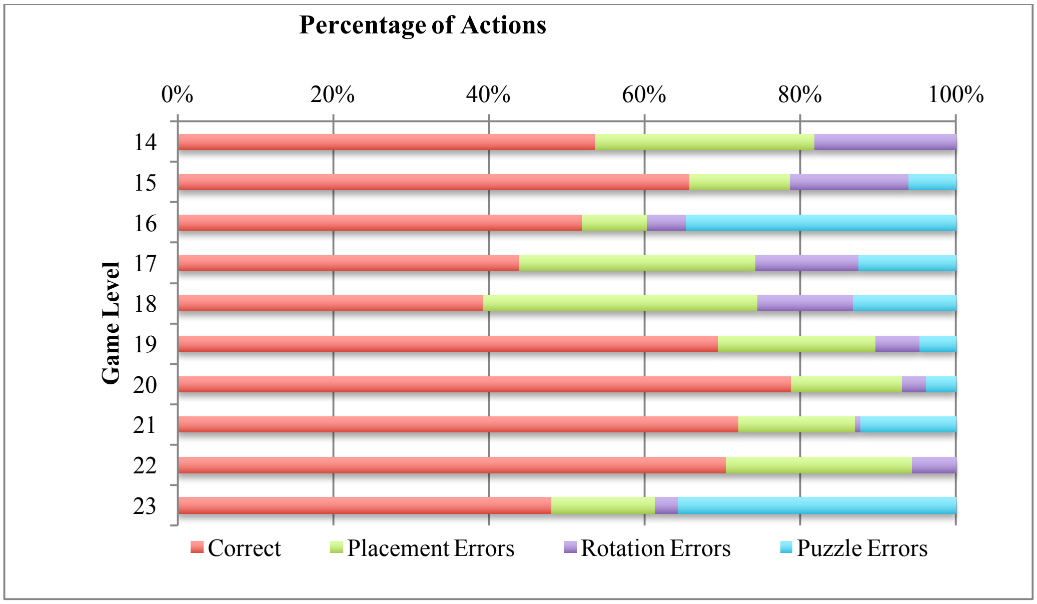


Figure 7: Error rates by puzzle level. Source: Eagle, et al. (2016)

There was considerable variability in the distribution of correct moves and errors across puzzle levels. These changes likely depend both on students' implicit science learning (i.e., they make fewer errors as their implicit understanding improves) and features of the puzzle design (e.g., number of obstacles, number of targets, novelty of puzzle). The next step is to validate the prevalence of science and puzzle errors against changes in students' pre/post assessments.

Research Questions & Hypotheses

The rest of this paper reports on the relationship between in-game measures of implicit science learning and changes in external pre/post assessment scores using HLM to account for the nesting of students within classrooms and teachers having multiple classes in the national implementation studies described above (Rowe, Asbell-Clarke, Bardar, and Edwards, 2017). To validate the GBLA detectors as assessments of implicit physics learning, we conducted a national implementation studies of both games, *Impulse* and *Quantum Spectre*. We compared changes in related physics understandings shown on pre- and post-assessment scores across classrooms assigned to one of three conditions:

1. **Bridge** classes where students played the game in and out of class, and the teacher bridged game-based learning to classroom physics learning using game examples and discussion;
2. **Game Only** classes where students played the games outside class and the teacher taught the relevant physics with typical instruction;
3. **Control** classes where students did not play the game and received typical instruction.

For both games, learners in Bridge classes—where learners played the games and had classroom instruction intended to bridge the implicit game-based learning to explicit classroom learning—showed significantly larger learning gains than students in classes that only played the game or had no access to the game (Rowe, Asbell-Clarke, Bardar, & Edwards, 2017). These findings show that teachers can effective bridge game-based learning to science learning on a class level, and leads to the ultimate research question of this paper: *How do in-game measures of implicit learning in* Impulse *and* Quantum Spectre *relate to external measures of implicit learning of the science phenomena (i.e., are they valid assessments)?*

We hypothesize that if the measures are valid, then players who exhibit game behaviors consistent with implicit science understandings will show larger gains on external learning assessments than players who exhibit those behaviors less often. Similarly, we hypothesize that players whose gameplay suggests a lack of implicit understanding will show smaller gains than players without those gameplay patterns. Table 2 summarizes how the salient implicit knowledge is hypothesized to map to the game strategy or move and how that maps to the evidence in the data logs for each salient phenomenon or construct.

Table 2: Connecting explicit and implicit science knowledge.

Explicit Science Construct	**Implicit Knowledge**	**Game Strategy**	**Game Behaviors/Moves**
Newton's First Law	Particles will stay in constant motion (speed and direction) until acted upon by a force.	Let particles not in danger float, use force to oppose incoming dangerous particles.	Player does not impart force (lets it float) or consistently clicks in the particle's path close enough to stop or slow it down.
Newton's Second Law	To change their motion heavier particles require more force than lighter particles.	Impart more force to move heavier particles than lighter particles.	Player consistently clicks more frequently next to heavier particles than lighter particles.
Law of Reflection	Angle of incidence equals angle of reflection.	Anticipate the path of the light beam for the various angles possible for lenses and mirrors.	Player consistently rotates mirrors to direct the laser beams correctly.
Slope	A linear path is determined by a consistent vertical change to horizontal change.	Anticipate the path of the light beam for the various grid placements of the optical devices.	Player consistently places optical device on the grid to direct the laser beams correctly.

Source: *Impulse*: Rowe, Asbell-Clarke & Baker (2015)

Applying these hypotheses to the in-game measures created for *Impulse* and *Quantum Spectre*, we anticipate that:

1. In-game measures of implicit understandings (Float, Oppose, Keeping the Player Path Clear strategies and correct differentiation of mass in *Impulse*) will be positively related to learning gains.
2. In-game measures of a lack of implicit understanding (incorrect differentiation of mass in *Impulse*, science errors in *Quantum Spectre*) will be negatively related to learning gains.

For the results presented here, these in-game measures were combined with hierarchical linear modeling (HLM) to relate individual students' implicit game-based learning to their performance on external science learning assessments.

An alternative hypothesis is that it is gameplay duration or achievement that predicts learning, regardless of specific gameplay behaviors. If this were true, players with greater durations of gameplay and who reached higher levels in the game should have greater gains between the pre and post assessments than players with shorter durations of gameplay or lower levels reached. To test this alternative hypothesis, we included the duration of game play and highest level reached in the HLM models.

Methods

Sample

This paper reports results from 17 *Impulse* teachers (19 Bridge and 19 Game Only classes) and 17 *Quantum Spectre* teachers (15 Bridge and 12 Game Only classes) drawn from the larger implementation studies are described in Rowe, Asbell-Clarke, Bardar, and Edwards (2017). The schools in these studies were located in 22 states.

Table 4: Teacher and student samples for the *Impulse* and *Quantum Spectre* gameplay analyses.

Study Group	Impulse				Quantum Spectre			
	# Teachers	# Classes	# Students	% Full Study	# Teachers	# Classes	# Students	% Full Study

Bridge	9	19	149	82%	8	15	161	67%
Game Only	8	19	150	72%	9	12	158	73%
Total	**17**	**38**	**299**	**77%**	**17**	**27**	**319**	**70%**

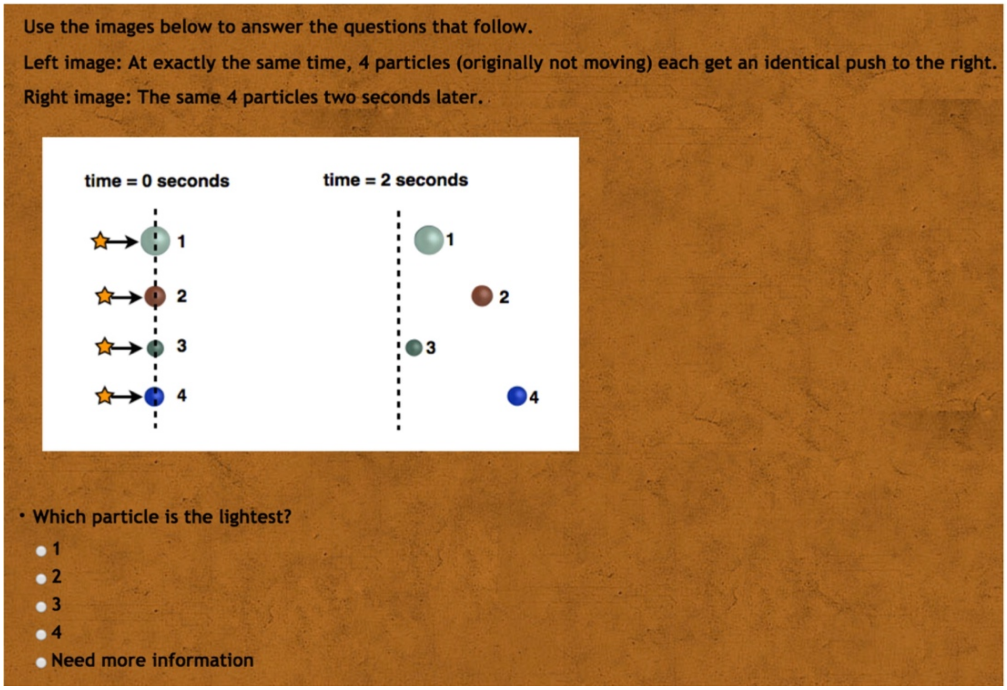
Data were not used from the earlier levels of either game where players were still learning the game mechanics. This excluded between 18% and 33% of the sample of each class, depending on the game and condition.

Measures

The analyses in this paper rely on three distinct data sources used in both implementation studies: pre/post assessments, teacher surveys, and game logs. All student assessment and gamelog data were collected through the game data collection architecture, *Data Arcade*.

Pre-Post Assessments. The pre/post physics learning assessments used with *Impulse* and *Quantum Spectre* were written to be answerable with an implicit understanding of the physics concepts (i.e., limited formalisms). Items were divided between pre- and post-assessments such that each assessment had similar but not identical items to avoid practice effects. Details about the development of these assessments can be found in Rowe et al. (2017). To ease interpretation of the HLM results, the pre- and post-assessments for both studies were standardized as Z-scores (subtract mean and divide by standard deviation) to have a mean of 0 and a standard deviation of 1. Thus, all HLM coefficients are reported in effect sizes (i.e., units are standard deviations of the post-assessment).

Impulse Assessments. Each assessment included six items, three dealing with Newton's First Law and three dealing with Newton's Second Law. For each topic, there was one question that resembled an animated version of a question from the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992; Thornton & Sokoloff, 1998), one question using an example from *Impulse*, and one using an excerpt from a NASA astronaut video. A sample item is presented in Figure 8.



Use the images below to answer the questions that follow.

Left image: At exactly the same time, 4 particles (originally not moving) each get an identical push to the right.

Right image: The same 4 particles two seconds later.

- Which particle is the lightest?
  - ○ 1
  - ○ 2
  - ○ 3
  - ○ 4
  - ○ Need more information

Figure 8: Sample Newton's First Law assessment item.

Both *Impulse* assessments had a maximum of 10 points possible (4 NFL, 6 NSL).[1] All students with perfect scores on the pre-assessments were dropped from these analyses to prevent ceiling effects—6% of the students in the full study. The pre- and post-assessment items had Cronbach's alpha of 0.48 and 0.58, respectively, indicating low internal consistency for the separate assessments. There are several potential explanations for these low values. The low values are likely due to NFL and NSL items being combined into a single scale because there were too few items for separate scales. The relatively small number of items and the skewing of the scores

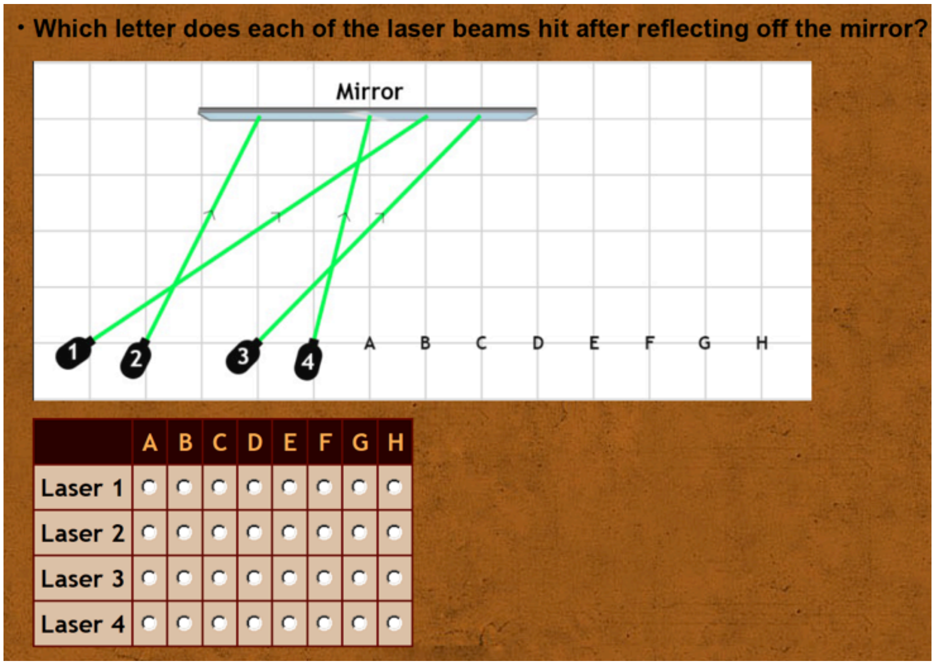
[1] One of the Newton's Second Law items was vaguely worded on the pre and post assessments. All responses were considered correct, so the minimum score on both assessments was 1. This item was not included in the Cronbach's alpha.

toward the high end of the scale also contribute to lower alphas. Finally, the internal consistency of students on the items before they have learned specific concepts on the pre-assessment could be lower due to random guessing or other unknown sources of error. When the pre- and post-assessment items were combined, Cronbach's alpha was 0.66.

Standardized scores were created using the mean and standard deviations. For the pre-assessment, the mean was 7.44 with a standard deviation of 1.53. The post-assessment had a mean of 8.25 and a standard deviation of 1.59. The within-group effect sizes (d_{Cohen}) ranged from 0.45 to 0.65. One-way ANOVAs showed significant differences in pre- and post-assessment scores by study group, with students in Bridge classes scoring significantly higher in both instances (Pre F (2, 496)=4.904, $p<0.05$; Post F (2, 496)=6.710, $p<0.05$).

Quantum Spectre Assessments. The *Quantum Spectre* assessments contained 12 pre and 13 post questions that required minimal formalisms to complete. Figure 9 is a sample Law of Reflection item. To account for the different number of items on the pre and post, we used the percentage of items answered correctly as the assessment score in these analyses. All students with perfect scores on the pre-assessments were dropped from these analyses—3% of the students in the full implementation study. These pre- and post-assessment items had good internal consistency (Cronbach's alpha for standardized items was 0.67 [pre] and 0.78 [post]).

The overall mean percentage correct and standard deviation were used to standardize the scores. Standardized scores were calculated using mean percentage correct (54% on the pre, 60% on the post) and standard deviation (22% on the pre, 24% on the post) for the entire group. Within group effect sizes ranged from 0.18 to 0.28. Students in Bridge classrooms answered significantly fewer questions correctly on both the pre- and post-assessment than students in Game Only classrooms (Pre: F (1, 388)=29.5, $p<0.01$; Post: F (1, 388) =16.2, $p<0.01$).



• Which letter does each of the laser beams hit after reflecting off the mirror?

	A	B	C	D	E	F	G	H
Laser 1	○	○	○	○	○	○	○	○
Laser 2	○	○	○	○	○	○	○	○
Laser 3	○	○	○	○	○	○	○	○
Laser 4	○	○	○	○	○	○	○	○

Figure 9: Sample Law of Reflection assessment item.

Teacher, classroom, and student characteristics. The analyses reported here focus on in-game measures of implicit learning, so only teachers with students playing the game in the Bridge and Game Only groups are included in these analyses. Demographic characteristics such as the socioeconomic status and percentage of minority students in the school community which prior research suggests are correlated with science learning (Wayne & Youngs, 2003; Dee, 2007; Sirin, 2005; Morgan et al., 2016; Reilly, Neumann, Andrews, 2015) were used in assigning teachers to these groups in the broader implementation study. With attrition, the distribution of teachers with these characteristics across the groups was not equal so these characteristics are included in these analyses as background variables.

Teacher surveys: School, Teacher & Classroom Characteristics. Common Core of Data were used to determine percentage of students receiving free/reduced price lunches and percent of students with minority status for public schools. Private school teachers reported similar information for their schools. All teachers were also asked their gender, the number of years they have taught science, and their degree of experience using educational games in their teaching. For each class in the study, teachers reported the number of students enrolled in the class, the percentage of students who completed the study, and whether or not the class was an Honors or Advanced Placement (college) level class. *Impulse* class sizes ranged from 8 to 28 while *Quantum Spectre* classes had 10 to 43 students. Of the 39 classes in each study, 11 *Impulse* (6 of 19 Bridge and 5 of 19 Game Only) and 8 *Quantum Spectre* (2 of 19 Bridge and 6 of 18 Game Only) were Honors/AP classes. Supplementary Tables 1 and 2 report the percentage of students in various types of schools and classrooms (https://edge.terc.edu/display/EDGE/Supplementary+Materials).

Data Arcade: Student characteristics. During the *Data Arcade* registration process, the following student-level characteristics were collected: gender, month of birth, and year of birth. Using the year of birth, students were classified as Freshman-Sophomores (for *Impulse*, born 1997 or later, for *Quantum Spectre* born 1998 or later) and Juniors-Seniors (for *Impulse* born 1996 or earlier, for *Quantum Spectre* born 1997 or earlier). While not a perfect classification due to students repeating grades, it is sufficiently granular for these analyses comparing groups of students on overall age trends. Supplementary Tables 3 and 4 report the percentage of students by gender, age, pre-assessment scores, game duration, and game achievement in the Bridge and Game Only classes for each game (https://edge.terc.edu/display/EDGE/Supplementary+Materials).

Measures of Implicit Physics Learning in *Impulse:*

Student-level measures of gameplay: Newton's First Law. The detectors of each learning strategy were applied to the data from 149 Bridge students and 150 Game Only students. As binary classification algorithms (Witten & Frank, 2005), the detectors of learning strategies not only produce an inference of whether a strategy is present or absent, but also produce a confidence in that inference. For example, if the Float detector has a confidence of 62% for a student action, there is a 62% probability that the student was using the Float strategy in that action. As Baker (2015) recommends, we average across confidences rather than transforming confidences into binary values. For example, if the detector indicated that a student had, in five actions, a 62%, 68%, 96%, 42%, and 92% probability of using the Float strategy, the average confidence for that student would be 72%. As such, the most likely estimate for how often this hypothetical student was using the Float strategy is 72% of their actions. For the HLM analyses, the average confidences for each strategic move were standardized. Table 5 reports the mean confidence and standard deviations for each strategic move used for standardization. There were no significant differences between in strategic move use by students in Bridge and Game Only groups.

Table 5: Mean Strategic Move Confidence and Standard Deviations.

Strategic Move	**Mean (S.D.)** **(n=299)**
Float	0.42 (0.08)
Toward Goal	0.38 (0.06)
Stop/Slow Down	0.09 (0.03)
Keep Player Path Clear	0.17 (0.06)
Keep Goal Clear	0.12 (0.02)
Buffer	0.06 (0.03)

Student-level measures of gameplay: Newton's Second Law. The 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). To do this we:

A. Identified the particle closest to the player's click.

B. Identified sequences where the player clicked repeatedly next to the same particle.

C. Identified set of sequences where the clicks were no more than 4 seconds apart

D. Measured the length of each identified sequence

E. Computed the average and standard deviation sequence length for each student's treatment of each colored particle.

With these average and standard deviation values (and the number of sequences for each student and color particle), we conducted a set of two-sample t-tests with pooled variance to determine whether students were differentiating between colors. A student was considered to be differentiating between colors if the means were statistically significantly different between two colors (taking the standard deviations and sample sizes into account). Because we were running multiple statistical tests for a substantial number of students, we used Benjamini and Hochberg's (1995) post-hoc method, as above.

Once we could identify which students were differentiating masses of the particles, we sorted them by whether those differences reflected differentiating mass correctly OR incorrectly. It was possible, for example, for a player to differentiate the masses of the Red and Blue particles correctly but not the White and Light Grey particles. Too few players reached the game levels with all 4 colors to test each color (mass) combination directly. For the HLM analyses, if a player had any significant differences between two colored particles, we recorded it as evidence of correct or incorrect differentiation. Two thirds of players (68%) had no evidence of correct or incorrect differentiation, 21% had evidence of correct differentiation only, 5% had evidence of incorrect differentiation only, and the remaining 6% had evidence of correct and incorrect differentiation.

Measures of Implicit Physics Learning in *Quantum Spectre*

Student-level measures of gameplay. These measures relied on the automated coding of Interaction Networks to identify each game move (action) as a correct move or as a placement, rotation, or puzzle error. To allow for the fact that students (a) used varying numbers of moves to solve the puzzles and (b) not all students completed Levels 14 through 23, the percentage of the total number of moves that were correct, placement errors, rotation errors, and puzzle errors were calculated. The mean error rate across all students was 19% placement errors (S.D. =11%), 7% rotation errors (S.D.=6%), and 12% puzzle errors (S.D.=7%). In the HLM models, we used standardized (z-scores) error rates.

Multilevel analysis

Hierarchical linear models were chosen to account for any common variance in post-assessment scores due to the clustering of students within classrooms. Using the SPSS MIXED linear models procedure, we estimated unconditional 2-level models with students nested within classrooms, using Restricted Maximum Likelihood (REML) and unstructured covariances. In unconditional 2-level models, a statistically significant percentages of the variance for was attributable to classroom level variation for *Impulse* (15 percent) and *Quantum Spectre* (36 percent).

Sets of covariates were added to the unconditional 2-level models in this order:

Set 1. Pre-assessment score (standardized)

Set 2. Study Group (Bridge, Game Only, Control)

Set 3. Student gender and age (Freshman/Sophomore versus Junior/Senior)

Set 4. Classroom Level Characteristics: Whether or not they were enrolled in class in which more than half of the students completed the study (1=Yes); whether or not they were enrolled in an AP/Honors science class (1=Yes); Class duration greater than 9 days (1=Yes, *Impulse* only); Class duration greater than 12 days (1=Yes, *Quantum Spectre* only).

Set 5. Teacher-Level Characteristics: Whether or not more than half of the students in the school received free-reduced price lunch (1=Yes); whether or not half of the students in the school were from minority groups (1=Yes); whether or not it was a Private school (1=Yes); whether or not the teacher was female (1=Yes); whether or not the teacher had more than five years of science teaching experience (1=Yes); and whether or not the teacher had prior experience using games in the classroom (1=Yes).

Set 6. In-game measures of (lack of) implicit science understanding (*Impulse*: Buffer, Stop/Slow Down, Float, Correct Differentiation of Mass, Incorrect Differentiation of Mass; *Quantum Spectre*: Percentage of moves that were Science Errors, Percentage of moves that were Puzzle Errors).

Set 7. Gameplay duration > 1 hour (1=Yes) and highest level completed (*Impulse*: Level 25 or higher versus not; *Quantum Spectre*: Level 22 or higher versus not)

Set 8. Interactions between Study Group and measures in Sets 2–7.

Study group and other statistically significant covariates were retained in the best fitting HLM models presented in this paper.

Results

For both games, HLM models with the in-game measures of implicit understanding provided a significantly better fit than models without them. Some, but not all, of the in-game measures were significantly correlated with improvements in external post-assessments after accounting for pre-assessment scores and teacher, classroom, and student characteristics. Gameplay duration and achievement was not significantly related to changes in either outcome.

Impulse gameplay as a valid measure of implicit science learning

The model with the in-game measures of implicit understanding of Newton's Second Law (Log likelihood=779.9; AIC=784; BIC=791) was a significantly better fit than the model (Log likelihood=789.4; AIC=793; BIC=801) without those measures (X^2 [2, N=299]=9.6, p<0.01). The model with the sole significant interaction (Log likelihood=773.3; AIC=777; BIC=785) improved the fit of the model (X^2 [1, N=299]=6.60, p<0.01). The best-fitting HLM model, which accounts for 33 percent of the variation at the classroom level, is presented in Table 7.

Table 7: Best fitting HLM model of estimated fixed effects on standardized *Impulse* post-assessment scores with in-game measures of implicit science learning.

					95% Confidence Interval	
Parameter	**Estimate**	**Std Error**	**df**	**Sig.**	**Lower**	**Upper**
Intercept	-0.03	0.19	183	0.88	-0.40	0.34
Pre-Score (Standardized)	0.22	0.06	291	0.00	0.11	0.33
Bridge classroom (vs. Game Only)	0.39	0.14	32	0.01	0.11	0.68
Female Student (1=Yes)	-0.39	0.10	290	0.00	-0.59	-0.19
Half of students complete study (1=Yes)	1.54	0.53	286	0.00	0.50	2.58
Any Correct Differentiation of Mass	0.25	0.12	289	0.03	0.02	0.48
Any Incorrect Differentiation of Mass	-0.32	0.17	280	0.05	-0.65	0.00
Half of students complete study* Any Incorrect Differentiation (1=Yes)	-1.32	0.54	287	0.01	-2.36	-0.26

The intercept coefficient represents the estimated outcome for male students who scored at the mean level of the pre-assessment, were in Game Only classrooms with less than half of the student participating in the study, and showed no evidence of correct or incorrect differentiation of mass. These students would score 0.03 standard deviations below the mean post-assessment score. The Pre-Score coefficient reflects the change in number of standard deviations of the post-assessment for every increase of 1 standard deviation on the pre-assessment. For every standard deviation increase on the pre-assessment, students would be expected to score 0.22 standard deviations higher on the post-assessment. Even after accounting for study group and pre-assessment scores, female students scored 0.39 standard deviations lower than male students on the post-assessment. Students in classes where more than half of the students completed the study scored 1.54 standard deviations higher on the post-assessment than students in classes where less than half of students completed the study.

Students who correctly differentiated particle masses by adjusting the number of clicks they made next to each particle type had post-scores 0.25 standard deviations higher than students who did not differentiate correctly. The opposite was also true—students who differentiated masses incorrectly (e.g., clicking more frequently next to lighter particles than heavier particles) had post-scores 0.32 standard deviations lower than students who did not exhibit that gameplay pattern. This pattern was moderated by whether or not students were in classes with the majority of students participating in the study.

Incorrect mass differentiation had a greater negative impact among students in classes with less than half of the students participating in the study. This could be because classes with more students in the study are more likely or able to dampen mistakes encountered in game-based learning.

***Quantum Spectre* gameplay as a valid measure of implicit science learning**

For the best-fitting *Quantum Spectre* model, Sets 3, 5, and 7 had no significant results, meaning student gender, teacher/school characteristics, gameplay duration, and highest level reached were not significantly related to changes in the percentage correct on the pre- and post-assessments. The model with the in-game measures of implicit understanding of slope and the Law of Reflection (Log likelihood=720.5; AIC=725; BIC=732) were a significantly better fit than the model (Log likelihood=728.2; AIC=732; BIC=740) without those measures (X^2 [2, N=319] = 7.7, p<0.05). The best-fitting HLM model (Log likelihood=715.4; AIC=719; BIC=727), which accounts for 74 percent of the variation at the classroom level, is presented in Table 8. This model with a significant interaction between study group and class duration was a marginally significantly better fit than the model without the interaction (X^2 [1, N=319] = 5.13, p<0.10).

Table 8: Best fitting HLM model of estimated fixed effects on standardized *Quantum Spectre* post-assessment score with in-game measures of implicit science learning.

					95% Confidence Interval	
Parameter	**Est.**	**Std Err**	**df**	**Sig.**	**Lower**	**Upper**
Intercept	0.15	0.17	13	0.42	-0.23	0.52
Pre-Assessment[1]	0.45	0.05	310	0.00	0.35	0.55
Bridge (vs. Game Only)	0.67	0.30	21	0.04	0.05	1.29
Honors/AP (1=Yes)	0.42	0.16	25	0.02	0.08	0.76
Greater than 12 days (1=Yes)	0.22	0.20	15	0.29	-0.21	0.65
%Placement Errors[1]	-0.08	0.04	297	0.07	-0.17	0.01
%Rotation Errors[1]	-0.13	0.05	307	0.01	-0.22	-0.04
Bridge*Greater than 12 Days (1=Yes)	0.80	0.33	21	0.02	0.12	1.48

[1] Standardized

Overall, after accounting for students' performance on the pre-assessment, students who exhibited more Placement and Rotation errors while playing the game performed more poorly on the post than students with lower science error rates. There was no significant impact of puzzle rates on post-assessment scores. The intercept reflects students in non-Honors/AP Game Only classes that spent fewer than 12 days of science instruction on optics; who had the mean percentage correct on the pre-assessment; and who had mean rates of placement errors and rotation errors. Those students scored 0.15 standard deviations above the mean post-assessment score. For every standard deviation increase in their pre-assessment score, students scored 0.45 standard deviations higher on the post-assessment. Students in Bridge classes scored 0.67 standard deviations higher on the post-assessment than students in Game Only classes. Students in Honors/AP classes scored 0.42 standard deviations higher on the post-assessment than students in non-Honors/AP classes.

For every standard deviation increase in their placement error rate, students lowered their post-assessment score by 0.08 standard deviations. Similarly, for every standard deviation increase in their rotation error rate, students decreased their post-assessment score by 0.13 standard deviations. Students in classes that spent greater than 12 days on science instruction around optics scored 0.22 standard deviations higher on the post-assessment than students in classes with less science instruction, but this effect is moderated by whether or not it was a Bridge or Game Only class. Students in Bridge classes who spent more than 12 days on optics instruction scored 0.8 standard deviations higher than all classes that spent less time. There was no significant difference between Bridge and Game Only classes with more than 12 days of optics instruction.

Limitations of the Study

There are several study limitations to bear in mind when interpreting these findings. Most important is the low internal consistency of the *Impulse* assessments (Cronbach's alpha was 0.48 for the pre; 0.58 for the post). When the items were combined, the alpha was 0.66. This is likely due to the small number of items necessitating combining items measuring Newton's First and Second Laws in one scale. The sample in this study was skewed toward high scores on the pre-assessment which restricts the range of learning gains. Both of these limitations work against finding significant relationships between the external assessments and the in-game measures. Both Newton's Second Law measures were significantly related to gains in these measures, but the Newton's First Law items were not.

Second, this study used a quasi-experimental design with a convenience sample. As described in Rowe, Asbell-Clarke, Bardar, and Edwards (2017), we balanced teachers assigned to each study group in terms of the school, teacher, and classroom characteristics described in the Methods section (public vs. private school, % of students receiving free or reduced-price lunch, etc.). With only 10 teachers needed per group for sufficient statistical power, there were too few teachers to match them across all of the characteristics and then randomize the assignment. The goal was to have each study group have comparable demographics for a fair comparison between the three study groups. Of the 42 *Impulse* teachers assigned to groups, 23 teachers (55 percent) completed all requirements for the implementation study. Similarly, 27 of the 63 *Quantum Spectre* teachers (43 percent) completed the study requirements. While the groups were balanced demographically at the beginning of the study, this attrition led to unbalanced demographics seen in Supplementary Tables 1 and 2.[2] These imbalances are a concern for the generalizability of the findings to populations underrepresented in this sample.

The final limitation is the complexity of the EDM techniques—detectors and interaction networks—can make it difficult for other researchers to replicate the findings exactly.[3] Whether or not another classroom implementation study would produce the same results of course depends on study-specific factors such as coding quality and reliability. Results would also depend on comparability of samples, gameplay duration, and variability in classroom implementation as would a replication of any naturalistic study, not solely studies with complex EDM techniques. The fact that we have used an emergent approach to the development of stealth assessment to produce valid measures in two games, not just one, suggests the viability of paying close attention to natural gameplay and building assessment mechanics from those observations.

Discussion & Implications

[2] Supplementary materials can be found at https://edge.terc.edu/display/EDGE/Supplementary+Materials

[3] We are happy to share the video coding manual, classification algorithms, and coordinates for science and puzzle errors with other researchers.

The GBLA methods using EDM in the studies of *Impulse* and *Quantum Spectre* show evidence of valid assessment of implicit physics learning. Using these techniques to understand how players behave in simulated ideal physics contexts enables stealth assessment and may reveal knowledge to educators and designers even before learners can articulate that knowledge through text or other formalisms.

In physics, implicit knowledge may be particularly important for educators because what is taught and tested in school may be counter to a learner's everyday experience are different. Motivated by the misconception literature in Physics education (e.g., McCloskey, 1983a, 1983b; Minstrell, 1982), the game *Impulse*, was able to reveal when and how players adopted strategies consistent with implicit understanding of Newton's First and Second Laws of Motion.

Researchers used the detectors built to study *Impulse* gameplay and were able to show that players who treated heavier mass particles with lighter particles performed better on external pre/post assessments of the salient phenomena of Newton's first and second laws of motion. Students who did not correctly differentiate their game behaviors to match the mass of the particles performed significantly worse on the post assessments. There was no relationship between the strategic moves thought to reflect an implicit understanding of Newton's First Law and external pre-post assessments. This may be due to the strong correlation between detectors of Newton's first law and second law of motion. This is understandable because the first is actually a specific case of the second. Newton's second law states that the magnitude of the Force required to accelerate an object equals the mass of object times the resulting net acceleration (F= ma). Newton's first law is just the explication of this law when the Force is zero. In the absence of a net force, the acceleration is zero and so the object stays in constant motion (or at rest). The game-based evidence used to assess Newton's first law (e.g. the absence of an imparted force followed by an opposing force when necessary) were less clear than the evidence used to assess Newton's second law, which was simply tracking the amount of force they applied for the different mass particles. This directness of the evidence was apparent in both the higher reliability of the detectors for Newton's second law and the greater the tie to explicit learning shown on the pre/post assessments.

In *Quantum Spectre,* Interaction Networks analysis reduced the wide array of player solutions to a manageable set of solution types that researchers and designers could interpret and human code. The different types of non-correct solutions were grouped into puzzle errors where the player did not understand how to solve the puzzle, and science errors that are consistent with the player not understanding the salient phenomenon. The science errors were further separated into (a) rotation errors that are consistent with lack of implicit understanding of the law of reflection and (b) placement errors that are consistent with lack of implicit understanding of the slope. Researchers related the percentage of science errors made by each learner with the learners' gains external pre/post-assessment items. The percentage of science errors players made in a puzzle were negatively related to their pre/post gains on related items, while the percentage of puzzle errors players made no difference, suggesting that science errors in *Quantum Spectre* are predictive of related lack of implicit science learning.

In studies of both *Impulse* and *Quantum Spectre*, HLM studies showed that the in-game measures of implicit learning developed through the use of data mining technique were correlated with external learning measures. Pre/post measures designed to measure implicit knowledge were used for each game, and in both cases learning gains could be predicted by players' patterns of game behavior. Teachers' use of bridging activities to connect game-based learning to explicit science content was a factor in student learning, suggesting that the game prepares learners with implicit knowledge and that teachers are able to leverage that implicit knowledge for improved classroom learning through bridging. Also meaningful were the lack of relationships between other measures of game performance (duration and highest level) and pre/post assessment changes in both games, suggesting that it is *how* that player plays the game that is a unique measure of implicit understanding.

These findings suggest implicit learning can be measured through behaviors, what players *do*, as suggested by Cook and Brown (1999), and providing the "transfer in" that may be useful for the preparation for future learning as described by Bransford and Schwartz (1999). These emergent forms of learning assessment show promise to alleviate some of the constraints imposed by an ECD model of assessment that relies on the prediction of learning trajectories, a priori, by the designers. When game-based learning assessments rely on preselected events, there is little room left to observe emergent strategies that learners exhibit that were unpredicted (or which are hard to exactly specify) but may reveal implicit understanding of the salient phenomena.

The results of this research also suggest that it may be possible to deploy richer, data-driven learning assessments in the future, mitigating some of the limitations of the current generation of assessments. When learning can be identified through behaviors, rather than through textual or verbal representations, we may be able to better measure knowledge in the full diversity of learners. This may be particularly helpful for learners with ADHD, ASD, and dyslexia who often have a wealth of implicit knowledge that is not articulated on assessments because construct-irrelevant factors including language processing difficulties and executive function challenges routinely impact the performance of students with disabilities on assessment instruments (Haladyna & Downing, 2004). Game-based learning research, and the measurement of learning within games may be boosted by the current emergence of neurophysiological sensors, such as wireless dry-electrode EEG devices. The coming of age of these portable and relatively inexpensive devices is making the observation of cortical activity associated with learning much more practical (Chi et al., 2013).

Finally, it should be noted that implicit learning assessments are only helpful when they inform educators and designers on how best to leverage the implicit knowledge for explicit learning, in the classroom or elsewhere. Implicit game-based learning that remains unarticulated may produce better game-play, but has little educative value. On the other hand, when that implicit knowledge is revealed and bridged by a teacher, it can be a valuable resource for learning. Similarly, game designers could use players' patterns of play to predict learning and produce real-time adaptations of games keeping players in their optimal learning zone, similar to Vygotsky's Zone of Proximal Development (Vygotsky, 1978). These are all exciting new affordances of data-driven implicit learning assessments, and thus further the call for more research in the design and validation of these powerful new tools (Shute & Ramini, 2017).

References

Asbell-Clarke, J., Edwards, T., Larsen, J., Rowe, E., Sylvan, E., & Hewitt, J. (2012). Martian Boneyards: Scientific Inquiry in an MMO Game. *International Journal of Game-Based Learning, 2*(1), 52-76.

Baker, R.S. (2015) Big Data and Education. 2nd Edition. New York, NY: Teachers College, Columbia University. http://www.columbia.edu/~rsb2162/bigdataeducation.html.

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

Baker, R. S., Ocumpaugh, J., Gowda, S. M., Kamarainen, A. M., & Metcalf, S. J. (2014). Extending log-based affect detection to a multi-user virtual environment for science *User Modeling, Adaptation, and Personalization* (pp. 290-300): Springer International Publishing.

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

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.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of research in education, 24*, 61-100.

Ceci, S. J., & Liker, J. K. (1986). A day at the races: A study of IQ, expertise, and cognitive complexity. *Journal of Experimental Psychology: General, 115*(3), 255.

Chi, Y. M., Wang, Y. T., Wang, Y., Maier, C., Jung, T. P., & Cauwenberghs, G. (2012). Dry and Noncontact EEG Sensors for Mobile Brain #x2013;Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 20*, 228-235. doi:10.1109/TNSRE.2011.2174652

Clark, D. B., Nelson, B. C., Chang, H.-Y., Martinez-Garza, M., Slack, K., & D'Angelo, C. 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 & Education, 57*(3), 2178-2195.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Cook, S. D., & Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization science, 10*(4), 381-400.

Dee, T. S. (2007). Teachers and the Gender Gaps in Student Achievement. *The Journal of Human Resources*, *1*, 528-554.

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

Eagle, M. & Barnes, T. (2014, July). Exploring differences in problem solving with data-driven approach maps. In proceedings of *Educational Data Mining (EDM2014)*, London, UK, pp. 76-83.

Eagle, M., Hicks, D., Peddycord III, B., & Barnes, T. (2015, March). Exploring networks of problem-solving interactions. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 21-30). ACM.

Eagle, M., Rowe, E., Hicks, A., Brown, R., Barnes, T., Asbell-Clarke, J., & Edwards, T. (2015, October). Measuring implicit science learning using networks of player-game interactions. Presented at the annual ACM Symposium on Computer-Human Interaction in Play, London.

Entertainment Software Association. (2014). Essential facts about the computer and video game industry: Sales, demographic and usage data (Annual report).

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave/Macmillan.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy*. (2nd Edition). New York: Palgrave MacMillan.

GlassLab. (2014). Psychometric Considerations In Game-Based Assessment (pp. 160): Institute of Play.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Halverson, R., & Owen, V. E. (2014). Game-Based Assessment: An Integrated Model for Capturing Evidence of Learning in Play. *International Journal of Learning Technology* (Special Issue: Game-Based Learning).

Halverson, R., Wills, N., & Owen, E. (2012, June). CyberSTEM: Game-Based Learning Telemetry Model for Assessment. Paper presented at the 8th annual Games+Learning+Society (GLS) Conference, Madison, WI, USA.

Hamari, J., Shernoff, D. J., Rowe, E., Coller. B., & Asbell-Clarke, J., Edwards, T. (2015). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior, 54*, 170-179. DOI: 10.1016/j.chb.2015.07.045.

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.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The physics teacher, 30*, 141.

Hicks, A., Eagle, M., Rowe, E., Asbell-Clarke, J., Edwards, T., & Barnes, T. (2016, April). Using game analytics to evaluate puzzle design and level progression in a serious game. Paper presented at the 6th international Learning Analytics & Knowledge conference, Edinburgh, U.K.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Macmillan.

Lave, J., Murtaugh, M., & de la Rocha, O. (1984). The Dialectic of Arithmetic in Grocery Shopping. In B. Rogoff and J. Lave, Eds., *Everyday Cognition: It's Development in Social Context*. Cambridge, MA. Harvard University Press., pp. 67-94.

Madden, M., Lenhart, A., Duggan, M., Cortesi, S., Gasser, Urs. (March 2013). Teens and Technology 2013. *Pew Internet & American Life Project and Harvard's Berkman Society for Internet & Society*.

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

McCloskey, M. (1983b). Naive theories of motion. *Mental models*, 299-324.

McGonigal, J. (2011). *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. New York: Penguin Press.

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.

Morgan, P., Farkas, G., Hillemeier, M., & Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Education Researcher, 45*(1), 18-35.

National Research Council. (2011). *Report of a Workshop on the Pedagogical Aspects of Computational Thinking*. Washington, DC: The National Academies Press.

Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*: Cambridge University Press.

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.

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction, 29*, 128-140.

Polanyi, M. (1966). *The Tacit Dimension*. London: Routledge. (University of Chicago Press. ISBN 978-0-226-67298-4. 2009 reprint).

Qian, M. & Clark, K. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior, 63*, 50-58.

Reilly, David; Neumann, David L.; Andrews, Glenda (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, *107*(3), 645-662.

Revell, A. D., Wang, D., Wood, R., Morrow, C., Tempelman, H., Hamers, R. L., ... & DeWolf, F. (2013). Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. *Journal of antimicrobial chemotherapy*, *68*(6), 1406-1414.

Roediger III, H. L., Putnam, A. L., & Smith, M. A. (2011). 1 Ten Benefits of Testing and Their Applications to Educational Practice. *Psychology of Learning and Motivation-Advances in Research and Theory, 55*, 1.

Rowe, E., Asbell-Clarke, J. & Baker, R. (2015). Serious game analytics to measure implicit science learning. In C.S. Loh, Y. Sheng, & D. Ifenthaler *Serious Game Analytics: Methodologies for Performance Measurement, Assessment, and Improvement.* Springer Science+Business.

Rowe, E., Asbell-Clarke, J., Bardar, E., & Edwards, T. (2017). Bridging Science Learning From Digital Games to the Classroom. Manuscript submitted for publication.

Rowe, E., Asbell-Clarke, J., Eagle, M., Hicks, A., Barnes, T., Brown, R., & Edwards, T. (2016, June). Validating game-based measures of implicit science learning. Paper presented at the 9th international conference on Educational Data Mining in Raleigh, NC.

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 Seventh international conference on Educational Data Mining Society in London.

Rowe, E., Bardar, E., Asbell-Clarke, J., Shane-Simpson, C., & Roberts, S. (2016). Building Bridges: Teachers Leveraging Game-Based Implicit Science Learning in Physics Classrooms. In

D. Russell & J. Laffey (Eds). *Handbook of Research on Gaming Trends in P-12 Education*. IGI-Global.

Sao Pedro, M., Baker, R.S.J.d., & Gobert, J.D. (2012). *Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information.* Paper presented at the 20th International Conference on User Modeling, Adaptation and Personalizations (UMAP 2012).

Sao Pedro, M. A., Baker, R. S. J. d., Gobert, J. D., 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. & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*, 1-19.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. *Serious games: Mechanisms and effects, 2*, 295-321.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58-67.

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.

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Steinkuehler, C., & Duncan, S. (2008). Scientific Habits of Mind in Virtual Worlds. *Journal of Science Education and Technology, 17*(6), 530-543.

Squire, K. (2011). *Video Games and Learning: Teaching and Participatory Culture in the Digital Age*. New York, NY: Teachers College Press.

Sternberg, R. J. (1996). Myths, countermyths, and truths about intelligence. *Educational Researcher, 25*(2), 11-16.

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

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*(4), 338-352.

Wayne, A. & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73(1), 89-122.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.

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