CHAPTER 2

CONSISTENCY AND VALIDITY IN GAME-BASED STEALTH ASSESSMENT

Valerie J. Shute and Gregory R. Moore

IAP PROOF
© 2017

ABSTRACT

Assessments need to evolve to accurately measure the higher-order skills that people need to be successful in the 21st century. We believe that game-based stealth assessment is one way to modernize assessments to meet this need. Stealth assessment refers to unobtrusively embedding assessments directly and invisibly into a gaming environment. While ample real-time data is available on a player's interactions with a game, a primary challenge in using stealth assessment in games is taking this stream of data and making valid inferences about players' competencies that can be examined at various points in time (to see growth), and also at various grain sizes (for diagnostic purposes). We suggest that reliability and validity can be achieved by following the evidence-centered design process to create stealth assessments and by using Bayesian Networks to accumulate evidence from the game. Reliability can be demonstrated through a combination of factor analyses and correlations and validity can be demonstrated by comparing

Technology Enhanced Innovative Assessment:
Development, Modeling, and Scoring From an Interdisciplinary Perspective, pp. 31–51
Copyright © 2017 by Information Age Publishing
All rights of reproduction in any form reserved.

the stealth assessments to external measures of the same competency. Through these processes, we can be confident that stealth assessments are measuring their target competencies accurately.

CONSISTENCY AND VALIDITY IN GAME-BASED STEALTH ASSESSMENT

You can discover more about a person in an hour of play than in a year of conversation.

—Plato

In this chapter, we start by presenting the current state of assessment, and then argue for the use of games as a viable alternative to traditional assessment, illustrated within the context of a recent stealth assessment project. The research project involves the game *Physics Playground* (Shute & Ventura, 2013), which includes three stealth assessments running concurrently and invisibly in the game—measuring physics understanding, creativity, and persistence directly from gameplay data. The underlying models and mechanisms that comprise the stealth assessments are presented, along with examples of how evidence was identified and accumulated in the models. Finally, we describe how we established validity and consistency for the *Physics Playground* stealth assessments, and conclude with a discussion of future research in this area.

IAP PROOF © 2017

State of Assessment

Since the 1960s, large-scale achievement testing has played a dominant and consequential role in the assessment of student learning outcomes. This role has come with a predictable set of rules and by-products. For example, the rules have included designing test items manually, imposing rigorous content and statistical checks to ensure psychometric defensibility, administering in paper-and-pencil format, and rank ordering students. The by-products of this focus have included the systematic assessment of students and comparability of learning outcomes across grades, years, and jurisdictions, but also, in some cases, measuring lower-level (superficial) skills in reading, math, and science, and providing ineffective forms of feedback to teachers, students, and policy makers.

The science of large-scale achievement testing has grown to such a level of sophistication, imposing a gigantic "footprint" in the assessment of learning, that it has systemically dwarfed classroom assessments. Classroom assessments—unable to be developed with the technical infrastructure to ensure psychometric defensibility—have become little more than tools for

----------> AU: Update?

preparing students to perform on the large-scale tests that really matter (see Shute, Leighton, Jang, & Chu, in press). However, at least two conditions now exist that indicate that these rules and effects are no longer tenable, forcing testing specialists and psychometricians to make significant changes in how assessments are conceptualized, designed, administered, and interpreted (Mayrath, Clarke-Midura, Robinson, & Schraw, 2012; Shute & Becker, 2010).

The two conditions forcing us to rethink assessments include advances in the learning sciences and technology. First, advances in the learning sciences indicate that acquiring and demonstrating new knowledge and skills occurs within an environment or pedagogical context, which includes (a) learners with specific cognitive and noncognitive profiles, and (b) tools to promote and evaluate learning (Pellegrino, Chudowsky, & Glaser, 2001). Second, technology has dramatically changed the environments and processes by which students learn and communicate, teachers instruct, and assessments are designed and administered. Paper-and-pencil tests are slowly becoming a thing of the past as assessments are now increasingly being viewed as adaptive and delivered online, employing interactive tasks and simulations (e.g., Gierl & Haladyna, 2012). This wave of innovation, ushered in by advances in the learning sciences and technology, has revolutionized the science of assessment, permitting greater ecological validity and feedback to students related to the breadth and depth of knowledge and skills learned in-situ, including so-called 21st century skills (e.g., critical thinking, creativity, collaboration, and problem solving). That is, advances in technologies and their integration with assessment systems have allowed for the assessment of multidimensional learner characteristics (cognitive, metacognitive, and affective) using authentic digital tasks (e.g., games and simulations).

However, despite these advances, assessments of learning are still mainly traditional—using multiple choice and short answer format for items—which may be efficient for measuring declarative knowledge, but not very effective for measuring higher-order skills. Essays can be used to measure some higher-order skills, but they are expensive and time consuming to both take and grade. In addition, traditional assessments are divorced from learning and tend to take place within inauthentic contexts. In many (if not most) classrooms today, instruction is designed to have discrete learning and assessment portions, with students needing to demonstrate their knowledge skills separately from (a) when and where they learned them, and (b) how they apply the competencies. This separation is perpetuated because teachers typically engage in the following cycle: teach some content, stop, test for the content, and then repeat for each new unit of instruction.

IAP PROOF ©2017

It is no longer sufficient for students to acquire declarative knowledge and basic skills in math and English. Today's students need to develop higher-order skills, such as problem solving, persistence, creativity, and collaboration (Partnership for 21st Century Skills, 2008), and become lifelong learners. However, these skills are hard to measure (Shute & Wang, in press). As a result, there are few valid and reliable assessments for 21st century skills. This is one reason that schools are reluctant to embrace these skills. Old ways of testing, such as multiple choices tests, cannot accurately measure learning and succeeding in a complex world. Thus, we need to rethink assessment, which, in this chapter, means using games as assessment vehicles.

----------> AU: Update?

Games as Learning Environments

There is a convergence between the core elements of a good game and the characteristics of productive learning (Shute, Rieber, & Van Eck, 2011). Our thesis in this chapter is that (a) learning is at its best when it is active, goal-oriented, contextualized, and interesting (e.g., Bransford, Brown, & Cocking, 2000; Bruner, 1961); and (b) learning environments should thus be interactive, provide ongoing feedback, grab and hold attention, and have appropriate and adaptive levels of challenge—all features of good games. Gee (2003) has argued that the secret of a good game is not its 3D graphics and other bells and whistles, but its underlying architecture in which each level dances around the outer limits of the player's abilities such that it is hard enough to be just doable (see also Csikszentmihalyi, 1990, on flow theory). Along the same line, psychologists (e.g., Vygotsky, 1987) have long argued that the best instruction hovers at the boundary of a student's competence. Finally, both well-designed games and productive learning processes employ ongoing feedback as a major mechanism of play/learning support.

Well-designed games can be seen as vehicles for exposing players to intellectual activities. People who want to excel at something—from surgeons to artists—spend countless hours making intellectual effort and practicing their craft. There is considerable support in the literature, going back more than 100 years, that practice substantially improves knowledge and skills (e.g., Bryan & Harter, 1899; Ericsson, Krampe, & Tesch-Römer, 1993; Newell & Rosenbloom, 1981; Schneider & Shiffrin, 1977; Shute, Gawlick, & Gluck, 1998; Thorndike, 1898). But practice can be boring and frustrating, causing some learners to abandon their practice and, hence, learning. This is where the principles of game design come in: Good games can provide an engaging and authentic environment designed to keep practice meaningful and personally relevant. With simulated visualiza-

IAP PROOF ©2017

tion, authentic problem solving, and instant feedback, computer games can afford a realistic framework for experimentation and situated understanding, and thus act as rich primers for active learning (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; Gee, 2003; Squire, 2006). Furthermore, within-game learning support enables learners to do more advanced activities and to engage in more advanced thinking than they could without such help (Vygotsky, 1987). The tricky part about including learning support in games is providing support that does not disrupt engagement while learners are immersed in gameplay, and reinforcing emerging concepts and principles to deepen learning and engender transfer to other contexts.

Play is voluntary, intrinsically motivating, and involves active cognitive and/or physical engagement that allows for the freedom to fail (and recover) and experiment freely (Klopfer, Osterweil, & Salen, 2009; Pellegrini, 1995; Rieber, 1996). Unlike “free play,” a game is usually a contest of physical or mental skills and strengths, requiring the player to follow a specific set of rules to attain a goal (Hogle, 1996). Thus, well-designed games are highly engaging and are guided by design principles for both the interface and game mechanics (e.g., Desurvire, Caplan, & Toth, 2004; Fullerton, Swain, & Hoffman, 2008). That is, we can use salient game features (e.g., problem solving, adaptive challenges, and targeted feedback) to engender motivation, which in turn will support engagement and ultimately learning. Adaptive challenges and dynamic performance feedback in a game help to create an optimal environment for diverse players which will foster the sense of flow and potentially cultivate the *growth mindset* that engenders effort-driven, challenge-centered competency development (Dweck, 2006).

In short, well-designed games are engaging, which is an important prerequisite to learning. They are also ubiquitous, as approximately 97% of youth play video games (Lenhart et al., 2008). For these reasons, we believe that well designed games can act as transformative learning environments that support skill development and meaningful learning across a range of critical educational areas. Indeed, research suggests that games can improve a variety of learning outcomes when properly implemented (e.g., Wilson et al., 2009). Just as importantly though, games are also excellent vehicles for stealth assessment, described next.

STEALTH ASSESSMENT

Given the goal of using well-designed games to support learning in school settings and elsewhere, we need to ensure that the assessments are valid, reliable, and unobtrusive (to keep engagement intact). The output from the assessments, however, should be transparent. That is, players should be aware of how they are doing relative to important competencies at any

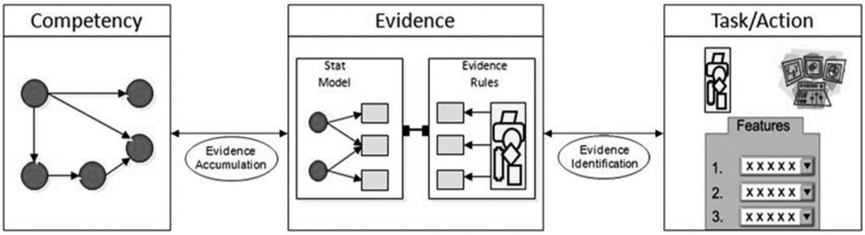
IAP PROOF © 2017

point in time to motivate learning. One way to meet these requirements is to use "stealth assessment" (Shute, 2011; Shute & Ventura, 2013). Stealth assessment refers to evidence centered design (ECD)-based models that are woven directly and invisibly into the fabric of the gaming environment (for more on ECD, see Mislevy, Steinberg, & Almond, 2003).

During game play, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess (e.g., problem solving, spatial skills, creativity, persistence). Evidence needed to assess the skills is thus provided by the players' interactions with the game itself (i.e., the processes of play). This evidence can, in turn, be contrasted with a summative score—the norm in educational environments. Making use of this stream of gameplay evidence to assess students' knowledge, skills, and understanding (as well as beliefs, feelings, and other states and traits) presents problems for traditional measurement models used in assessment. First, in traditional tests, the answer to each question is seen as an independent data point. In contrast, the individual actions within a sequence of events in a game are often highly dependent on one another. For example, what one does in a game at one point in time affects subsequent actions later on. Second, in traditional tests, questions are often designed to measure particular, individual pieces of knowledge or skills. Answering the question correctly is evidence that one may know a certain fact: one question—one fact. But by analyzing a *collection* of actions within gameplay (where each action provides incremental evidence about the current mastery of a specific fact, concept, or skill), stealth assessments can infer more accurately what learners know and do not know at any point in time. Now, because we typically want to assess a whole cluster of skills and abilities using evidence coming from learners' interactions within a game, methods for analyzing the sequence of behaviors to infer these abilities are not as obvious. As suggested above, stealth assessments that use ECD-based models can help address these problems.

Figure 2.1 shows the three main models of ECD: the competency, evidence, and task models. We first describe the process of assessment design. While assessment design is an iterative process, an assessment designer tends to work from left-to-right (i.e., from the competency model to the evidence model to the task model). The competency model defines what you want to be able to say about the learner (i.e., their knowledge, skills, and/or attributes). Development of the competency model involves identifying and structuring the relevant competency variables. The evidence model establishes the statistical relationships among the variables in the competency model and the observable variables obtained from game play. The evidence model also defines specific rules for automatically "scoring" obtained data. Finally, the task model defines the features of the tasks that will elicit the evidence that will ultimately inform the competency variables.

IAP PROOF © 2017



Source: Mislevy, Steinberg, and Almond (2003).

Figure 2.1. Three main models of ECD.

When using ECD to assess a learner's performance, one works in the opposite direction: from the task model to the evidence model to competency model. A learner's performance on a task provides a steam of evidence. Evidence rules pull relevant information from the stream of game play data, score it, and statistically link the scored, observed evidence to relevant competency variables. The competency model then updates the estimates of each competency variable in the model.

Stealth assessment (Shute, 2011) embeds the competency and evidence models created via ECD deeply into the learning/gaming environment such that the line between learning and assessment is blurred. This allows us to (a) extract dynamic, ongoing information of various grain sizes from the learner in real-time, (b) make accurate inferences of the learner's competencies at any time, and (c) support learning by reacting in immediate and helpful ways. Because stealth assessment is intended to provide support to the learner, it is mainly used for formative purposes. If the purpose of the assessment is summative, stealth assessment can still be used, although it would not tap the full potential diagnostic and support capabilities of stealth assessment.

To illustrate, the stealth assessment process (Figure 2.2) begins with a student playing a game. While she is playing the game, she is producing a dense stream of performance data (arrow 1). This performance data is captured in a log file, then analyzed and scored (arrow 2). The output of the analysis is sent to the student model (arrow 3), which uses the data to update its estimates of the student's competencies. The estimates of the student's competencies can then be used to provide feedback and other forms of learning support to the student during gameplay (arrow 4). This cycle repeats as long as the student is producing performance data. We will next discuss a recent project in which we applied stealth assessment to measure students' qualitative physics understanding, creativity, and persistence.

IAP PROOF © 2017

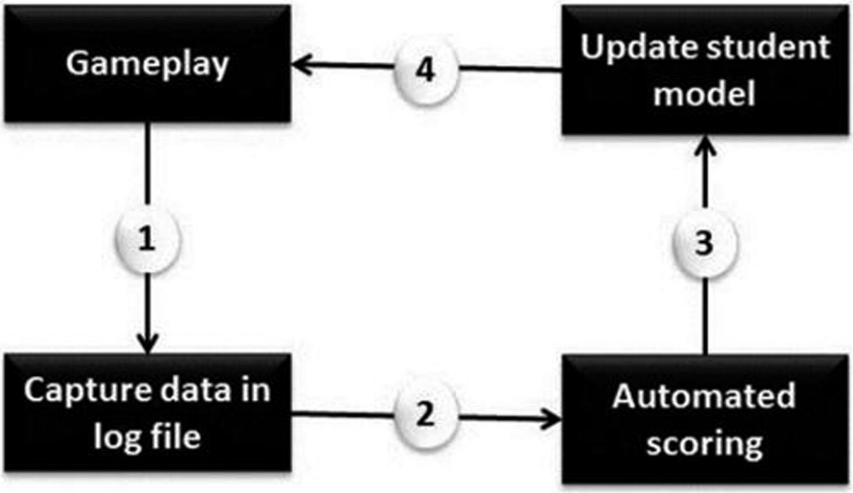


Figure 2.2. Stealth assessment process.

IAP PROOF ©2017

PRESENT WORK

We developed and applied stealth assessments in *Physics Playground* (formerly *Newton's Playground,* see Shute & Ventura, 2013). *Physics Playground* is a 2D physics-based computer game. The learning goal of this game is develop the player's understanding of qualitative physics within the context of simple machines (Shute, Ventura, & Kim, 2013). In the game, players need to guide a green ball to a red balloon. They do so by using their mouse to draw various simple machines (which are called "agents of force/ motion" in the game). These agents include ramps, levers, pendulums, and springboards (Table 1.1). All objects in the game follow the basic rules of physics (e.g., gravity, Newton's 3 laws). For example, Figure 2.3 shows a screenshot from one level in the game. The green ball is cradled and suspended in the middle of the screen (the purple and blue objects). To get the ball to the red balloon in the top left, the player has drawn a pendulum on the right (in red). When put into motion, the pendulum will hit the cradle, sending the cradle and the ball towards the balloon (as demonstrated by the arrows in blue).

AU: IAP does not print color images. Please change text as needed. <----------

There are 74 problems in the game, each of which has many possible solutions. Therefore, learners may play a level multiple times to solve the problem in different ways. However, some strategies are more effective than other. For example, in the problem shown in Figure 2.3, ramps are unlikely to be useful. We wanted to capture when players were using effective strategies and tools and demonstrating mastery over the game. Thus, each level

Table 2.1. Agents of Force/Motion Available in *Physics Playground*

Agent of Force/Motion	*Definition*
Ramp	Used to change the direction of the motion of the ball (or another object)
Lever	Rotates around a fixed point usually called a fulcrum or pivot point. Moves the ball vertically.
Pendulum	Directs an impulse tangent to its direction of motion. Secured at the top by a pin.
Springboard	Stores elastic potential energy from falling weight; becomes kinetic as weight is released. Moves the ball vertically.

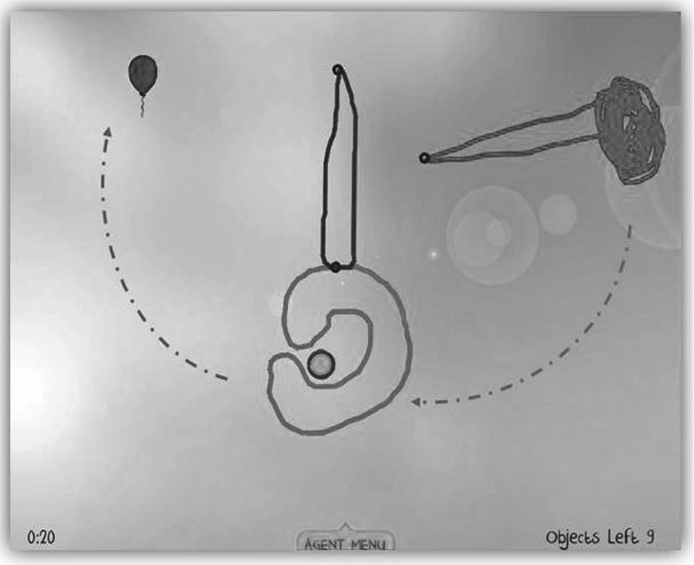


Figure 2.3. Screenshot from *Physics Playground.*

in the game has three possible levels of performance: not completed, a silver trophy, or a gold trophy. Players receive a silver trophy if they simply complete the level, but receive a gold trophy if they are able to complete the level under par (usually with three or fewer objects). Receiving a gold trophy suggests that the player has mastered (or is in the process of) mastering the competency.

To satisfy the requirements of the task model for our stealth assessments and ensure that all players are confronted with challenging levels, we designed levels with a wide range of difficulty. We used the following scheme to define each level's difficulty on a scale from 0 to 6:

- Relative location of ball to balloon: If the balloon is above the ball, the player is forced to use the agents of force/motion to impart a force and thus raise the height of the ball to solve the problem (0–1 point).
- Obstacles: If the pathway between the ball and the balloon is obstructed, the player must project the ball in a specific direction to solve the problem (0–2 points).
- Distinct agents of force/motion: More complicated problems require the player to use more than one agent of force/motion to solve the problem (0–1 points).
- Novelty: If the problem is unlike any other problem the player encountered, the player cannot easily determine the solution from prior experiences (0–2 points).

Based on these difficulty indices, the levels were organized into 7 groups of increasingly difficult problems. Each group, called “playgrounds” in the game, contained around 10 levels.

Project Findings

After playing *Physics Playground* for 4 hours with no instruction, students in the study demonstrated significant learning gains from pretest to posttest, $F(1, 153) = 4.24, p < .05$. Additionally, students tended to enjoy the game (1 = dislike; 5 = like; $M = 4$, $SD = 1$), with male and female students enjoying the game equally after controlling for the pretest. While these results are very promising, they do not use our stealth assessments to measure students’ competencies directly from gameplay in real-time, and are not our main focus. However, to understand the results from our stealth assessments, we first need to describe how we know they are valid and consistent. This involves explaining the underlying structure of the assessments, how evidence is identified and accumulated, and, most importantly, what procedures we used to test for validity and consistency. We discuss these issues in detail next.

MODELS FOR PHYSICS PLAYGROUND

Using ECD, we designed a competency model for each of our three focal constructs: qualitative physics understanding, creativity, and persistence. These three constructs were combined under a broader category of “Success in Physics Playground” (Figure 2.4). For the purposes of this chapter, we will describe the creation of the conceptual physics understanding competency model in depth. However, the process is similar for the other two competency models.

IAP PROOF ©2017

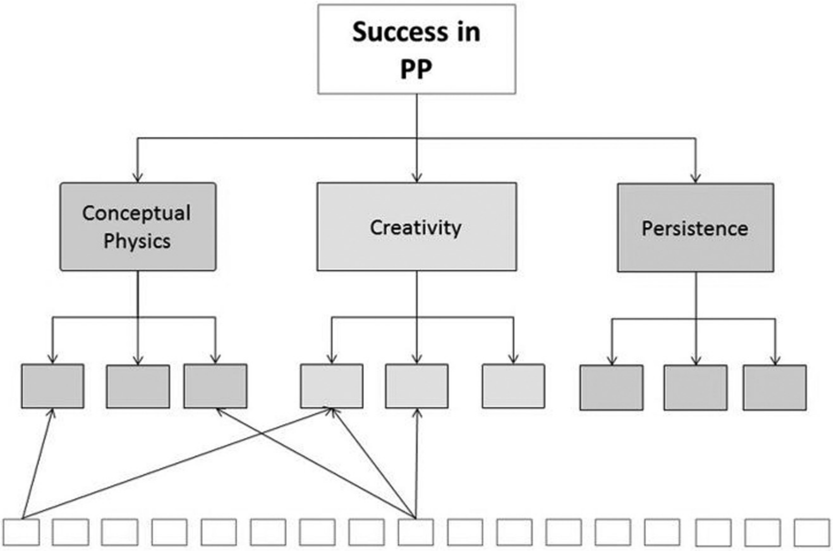


Figure 2.4. Simplified representation of the complete competency model.

IAP PROOF
© 2017

The competency models in Physics Playground were developed by conducting an extensive literature review for each of the three primary constructs. Based on the literature reviews, we then created graphical models, received expert feedback on the models, and subsequently refined them. To illustrate, for the physics competency model, we identified two broad categories of physics understanding from the literature: (a) Potential and Kinetic Energy, and (b) Angular Momentum.

Potential energy exists when a force acts on an object to restore the object to its resting point (or lower energy configuration). For example, when a springboard—like a player in *Physics Playground* could create—is bent downward, it exerts an upward force to return to its unbent position. The action of bending the springboard down stores energy in the springboard equal to the work done by the springboard in returning it to its resting point. When the bent springboard is released, the stored energy is converted into kinetic energy. The same holds true for the force of gravity. When an object is held at some height, it has gravitational potential energy that, when dropped, is gradually converted to kinetic energy. The angular momentum of an object about a point of reference is the product of its position and linear momentum. A useful example is a pendulum—like a player might create in the game. When it is released, a pendulum's angular momentum increases (and hence its speed) due to the unbalanced torque, which is due to the force of gravity.

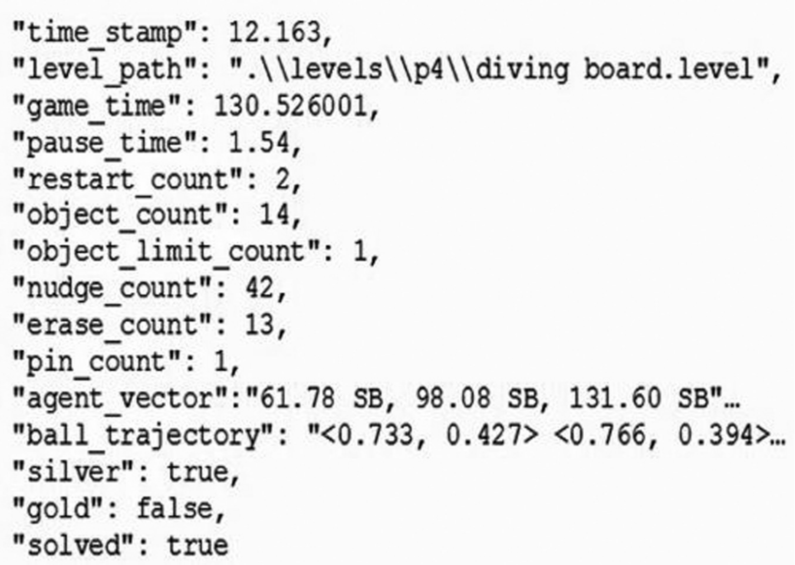
We then linked our four agents of force and motion to the two main physics constructs (see Figure 2.6 for a graphical representation). We also defined observable variables, or indicators, within *Physics Playground* that would provide evidence of effective creation and application of the agents and, by extension, the learner's physics understanding. All of the indicators we defined were equally applicable to all four agents. One indicator included trophy level data for a given level (i.e., gold, silver, or none), another reported the time, in seconds, spent on the level, and so on. Collectively, these indicators and their values across gameplay inform our estimates of the learner's knowledge about the agents, her understanding of the two main categories of physics, and her overall understanding of conceptual physics. Thus, if a player solves a level with a pendulum very quickly and using three or fewer objects with a pendulum, she likely has high pendulum knowledge, which suggests that she has a good understanding of angular momentum, which in turn suggests that she has a good overall understanding of conceptual physics. The specifics of how this type of evidence is identified and accumulated in real-time are described next.

Evidence Identification

IAP PROOF ©2017

To identify evidence and "score" gameplay performance, *Physics Playground* generates log files while students play the game. These log files capture a wide variety of gameplay behaviors that provide evidence related to our target competencies. Figure 2.5 displays a portion of one student's log file from a single level in the game. The metrics in this log file include, among others, the time spent on the level, the number of objects created, the number of restarts, whether the player received a silver/gold trophy, and the trajectory of the ball.

One particularly important piece of information is the agents of force/motion that the player creates during the solution process. Since players draw these agents with the mouse, and everyone draws differently, the game has a system that allows it to identify which agent the player intends based on the object he or she has drawn. The agent identification system is based on the idea that each agent has certain unique features that distinguish it from other agents. For example, pendulums rotate around a single pin, so any drawn object that does this is most likely a pendulum. Ramps tend to be in contact with the ball for a longer period of time. Levers tend to rotate when the ball comes into contact with them. Springboards tend to be connected to other objects with two pins and include a weight—either attached to the springboard or dropped on it and then deleted. In this way, the game is able to classify the player's drawn object with more than 95% accuracy (compared with human ratings) and provide detailed information



```
"time_stamp": 12.163,
"level_path": ".\\levels\\p4\\diving board.level",
"game_time": 130.526001,
"pause_time": 1.54,
"restart_count": 2,
"object_count": 14,
"object_limit_count": 1,
"nudge_count": 42,
"erase_count": 13,
"pin_count": 1,
"agent_vector":"61.78 SB, 98.08 SB, 131.60 SB"…
"ball_trajectory": "<0.733, 0.427> <0.766, 0.394>…
"silver": true,
"gold": false,
"solved": true
```

Figure 2.5. Sample log file.

IAP PRO
© 2017

on the agents drawn per level in the log files. For a more detailed explanation of this system, see Shute and Ventura (2013) and Shute, Ventura, and Kim (2013).

Evidence Accumulation

Once evidence is captured and analyzed in the log files, the game needs to accumulate this evidence to estimate the learner's competencies. This is accomplished by a script that sums the data across all instances in each level to produce raw indicator data per level (e.g., time on level, number of restarts per level). The data per level are then summed across all levels for some of the variables to create session level variables (e.g., number of levels attempted, applicable agents created). Finally, we use cut scores, obtained from the frequency distributions of the raw observables in our study, to define the Low, Medium, and High levels for each indicator.

The data are accumulated in Bayesian Networks (Figure 2.6), or Bayes nets, which use conditional probabilities to represent a learner's competencies. We created 74 Bayes nets for *Physics Playground*, one for each level in the game, because each level differs in terms of difficulty and discrimination parameters. These networks are implementations of the competency model and its associated indicators. To properly implement these models, though, we needed to define conditional probability tables for each of the nodes in the network, which represent the competency variables (in white) and the indicators (in green). The conditional probability tables (CPTs) were initially based on student pilot data.

Once we defined the prior probabilities for each CPT, we could then use the Bayes nets to estimate the unobserved variables of each player. For example, the Bayes net fragment in Figure 2.7 is updated with two observables. The player solved the level twice using two different agents. She received a silver trophy using a *lever* and a gold trophy using a *pendulum*. These two observables update the Lever Knowledge and Pendulum Knowledge nodes respectively. These competency nodes influence the estimates of the player's knowledge of Potential/Kinetic Energy and Conservation of Angular Momentum, and ultimately Newton's Three Laws. In turn, the estimates of the student's knowledge of Potential/Kinetic Energy, Angular Momentum, and Newton's 3 Laws influence the estimates of Ramp Knowledge, Springboard Knowledge, and the likelihoods of getting specific trophies using a ramp or springboard in the level.

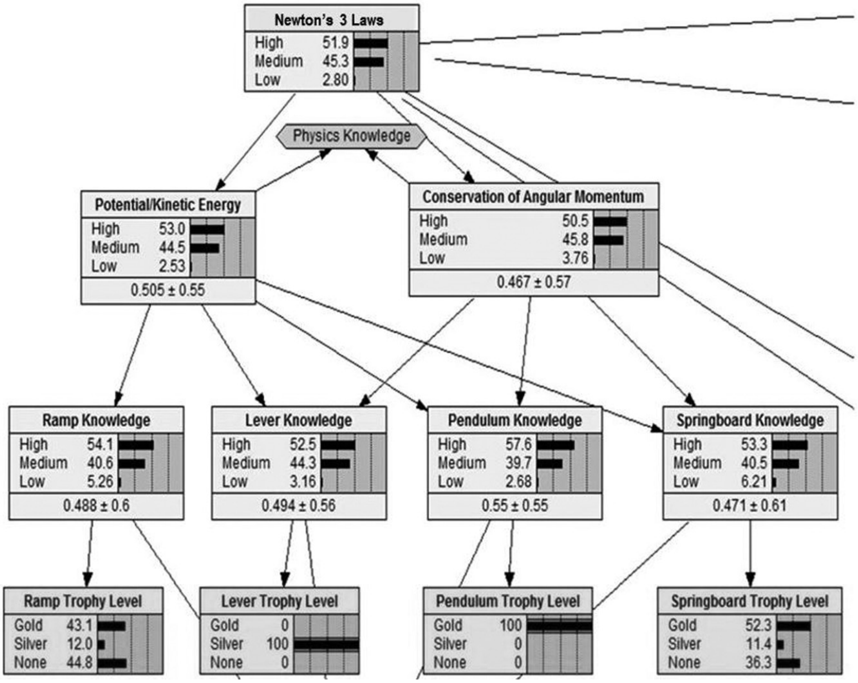


Figure 2.6. A Bayes net fragment instantiated with indicator data.

Based on this illustration, the player is currently estimated to be either at a medium or high state relative to most of the competency variables. For cases where the estimated values between states are quite close (e.g., Conservation of Angular Momentum, *p*(high) = .51 and *p*(medium) = .46), we derived a rule to decide if more evidence is needed for that node. After calculating the three absolute values between all pair-wise states, if any two of the three values are less than or equal to .15, then more evidence is needed. If not, we can calculate Expected a Posteriori (EAP) values to reduce the triplet of values to a single value ranging from –1 to 1. This is defined as *p*(high) –*p*(low). For reporting purposes, the student will have mastered the competency if EAP falls in [0.34, 1]; nearly mastered the competency if EAP falls in [–0.34, 0.33]; and (c) not mastered the competency, if EAP falls in [–1, –0.33]. These probabilistic models of players' knowledge are not useful if they do not meet the basic psychometric properties of consistency and validity, discussed next.

CONSISTENCY AND VALIDITY OF STEALTH ASSESSMENTS

Consistency

The learners were allowed to play any of the 74 levels they wanted and, therefore, not all learners attempted the same problems. This made it difficult to examine the reliability (i.e., consistency) of the stealth assessments simply by calculating Cronbach's alpha on the levels. Instead, we used a variety of other analyses and metrics to explore the consistency of our stealth assessments. For the purposes of this chapter, we will only describe how we examined the consistency of our stealth assessment for the conceptual physics competency. The process was the same for the stealth assessments of the creativity and persistence competencies.

First, we conducted a confirmatory factor analysis on the gold trophy (i.e., mastery) data to examine the consistency of the constructs (Figure 2.7). The four factors were highly intercorrelated and the CFA suggests that the mastery data (gold trophies) per agent fit a single factor well and have a small error variance. We also calculated the intraclass correlation on the gold trophy data and found a high correlation between each of the four agents of force/motion (r = 0.85).

Additionally, we examined the consistency among gold trophy performance for varying levels of difficulty: easy (levels in Playgrounds 2 and 3), medium (Playgrounds 4 and 5), and hard (Playgrounds 6 and 7). As expected, gold trophy performance was highly correlated between easy, medium, and hard levels, with an intraclass correlation of r = 0.82. The individual correlations between the three difficulty levels were also significant: easy-medium, r = 0.77; easy-hard, r = 0.53; medium-hard, r = 0.66.

IAP PROOF © 2017

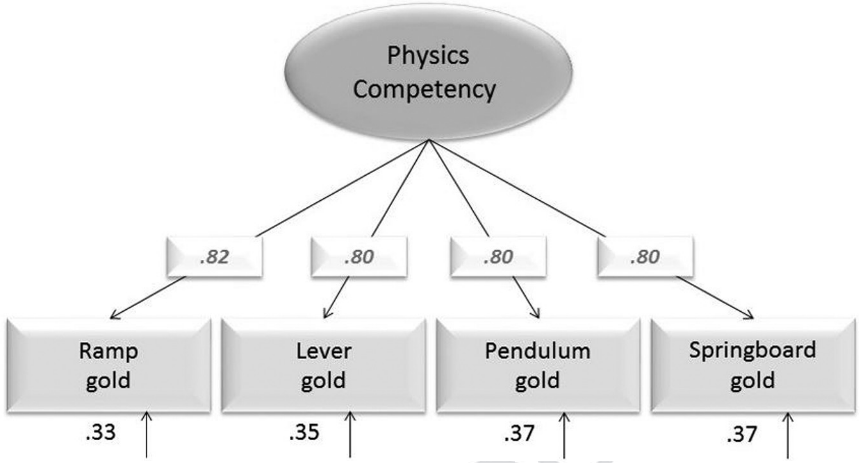


Figure 2.7. Confirmatory factor analysis for mastery with each agent of force/motion.

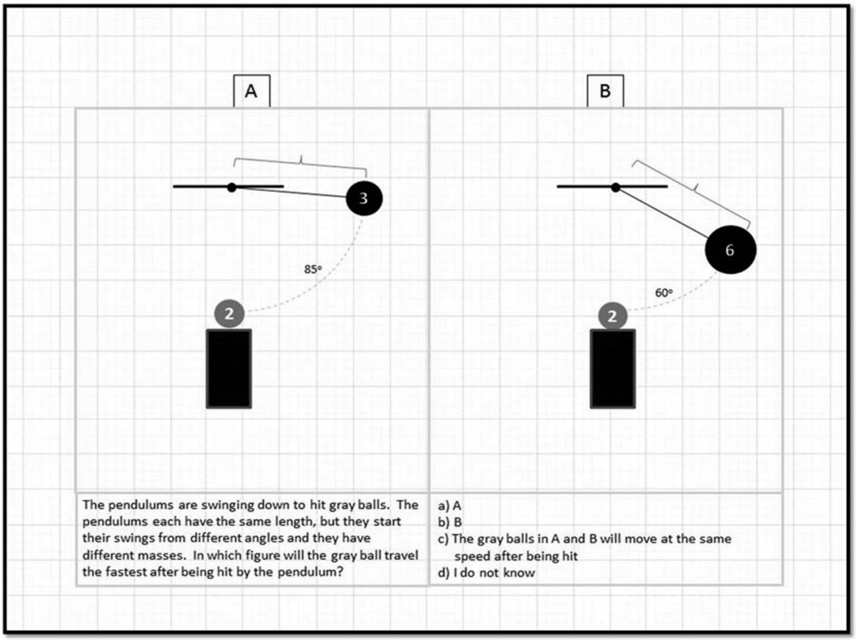
Finally, we computed Cronbach's alpha for gold trophy performance on levels that the majority of participants attempted. Most of these levels were of easy or medium difficulty, though some hard problems were also included. We first identified problems that were solved by more than 100 students. This produced a list of 29 levels (out of 74) that were solved by 110 students (out of 168). We then calculated a Cronbach's α on the performance data for the 29 levels. This analysis suggested that there is consistency among gold trophy performance in the levels, α = 0.87. Thus, the CFA, intraclass correlations, and alpha values all suggest that our stealth assessment for qualitative physics knowledge is consistent for gold trophy performance across agents of force/motion and varying difficulty levels. However, this does not necessarily mean that our assessment is actually measuring qualitative physics knowledge. To determine that, we need to examine the validity of the stealth assessment.

IAP PROOF © 2017

Validity

To validate the stealth assessments, we compared them to validated external measures of the same competencies. In general, we expected that the correlations between each stealth assessment and the corresponding external tests would be reasonably high, but not too high. That is, the external measures were "standard" measures, meaning that they were fairly limited and not completely comparable to the stealth assessments. We tested for both convergent validity (i.e., the stealth assessment and external measure of the same construct were significantly correlated), and divergent validity (i.e., the stealth assessment and external measure of a different construct were not correlated).

We describe the validation studies for persistence and validity elsewhere (see Ventura & Shute, 2013; Shute & Ventura, 2013). For the purposes of this chapter, we describe the validation of the physics stealth assessment. Towards that end, we developed, with two physics experts, a conceptual physics external test. The test contained 24-items consisting of two isomorphic forms (12 items each), informed by the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992) and the Mechanics Baseline Test (Hestenes & Wells 1992). Learners completed both forms—one as a pretest and the other as a posttest (in a counterbalanced design). The test contained both constructed responses and multiple choice items (Figure 2.8).



The pendulums are swinging down to hit gray balls. The pendulums each have the same length, but they start their swings from different angles and they have different masses. In which figure will the gray ball travel the fastest after being hit by the pendulum?

a) A
b) B
c) The gray balls in A and B will move at the same speed after being hit
d) I do not know

Figure 2.8. Example item from the conceptual physics external measure.

The pretest scores from the external measure of physics knowledge were significantly correlated with the number gold trophies (i.e., the stealth assessment measures of mastery) for each agent of force/motion, with the more challenging agents (pendulums and springboards) being more strongly correlated: Ramp ($r = 0.27, p < 0.01$), Lever ($r = 0.22, p < 0.01$), Pendulum ($r = 0.31$, $p < 0.01$), Springboard ($r = 0.40$, $p < 0.01$). The pretest scores from the external measure were also significantly correlated with the estimated overall physics competency ($r = 0.41, p < 0.01$), using the expected a posteriori (*EAP*) value per student. Finally, the pretest and posttest scores of our external physics measure were significantly correlated, $r = 0.60, p < 0.01$. These results meet our expectations and suggest that our conceptual physics stealth assessment is valid. Additionally, the results support our use of gold trophies as measures of mastery in the game. While these findings are positive, there is still more work to do to further validate the stealth assessments. For instance, in the future, we plan to compare the stealth assessments to various other external measures of conceptual physics knowledge to gauge validity and transfer.

IAP PROOF ©2017

CONCLUSION

In order to prepare students for success in the 21st century, we need to support the development of higher-order competencies, such as persistence, problem solving, and creativity. Supporting these complex competencies requires that we design innovative assessments that are valid and reliable. Using stealth assessments embedded in gaming environments is one potential way to address this need. Well-designed games are engaging and enable learning within complex, realistic, and relevant environments and evidence-centered design is a process that enables the development of valid assessments. Stealth assessment is a specific implementation of evidence-centered design that allows us to gather evidence from game players in real-time, without disrupting engagement, and use that information to support learning.

The present work suggests that we can gather good evidence for accurate, real-time estimates of competencies using stealth assessment. The information that is gathered from these assessments can be used for a variety of purposes. Teachers can use the competency estimates to adjust their instruction and provide meaningful feedback on a student's strengths and weaknesses. Students can use the competency estimates to gauge and reflect on their progress. Games and other computer-based environments can use the competency estimates to select new experiences and problems to present to players.

We are expanding on this work by building stealth assessments for various cognitive and noncognitive personal attributes (e.g., problem solving skill, spatial ability, affective states, engagement), as well as knowledge and skill acquisition across various domains (e.g., mathematics and computer programming). Additionally, because students in the present study demonstrated significant learning gains, we are also continuing our work with *Physics Playground*. Specifically, we will be adding learning supports to the game (i.e., targeted feedback, curriculum) to help students move from an informal understanding of physics to a formal understanding.

Through stealth assessments, we hope to be able to accurately assess a variety of learner competencies in authentic, engaging learning environments. Accurate assessments will lead to a better understanding of the current states and disposition of students, which will allow us to design instruction and other interventions to help students meet their goals. With proper help, students will be able to achieve high learning outcomes and obtain the skills that they will need to be successful. For these reasons, we view stealth assessment as an important step towards preparing learners for the 21st century.

IAP PROOF © 2017

REFERENCES

Barab, S. A., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development, 53*(1), 86–107.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). Washington, DC: National Academies Press.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21–32.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review, 6*, 345–375.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.

Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems,* 1509–1512. Vienna, Austria.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York: Random House.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Fullerton, T., Swain, C., & Hoffman, S. (2008). Game design workshop: A playcentric approach to creating innovative games (2nd ed.). Boca Raton, FL: Elsevier Morgan Kaufmann.

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

Gierl, M. J., & Haladyna, T. M. (2012). Automatic item generation: An introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 3–12). New York: Routledge.

Hestenes, D., & Wells, M. (1992). A Mechanics Baseline Test. *Physics Teacher, 30*(3), 159–167.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *Physics Teacher, 30*, 141–151.

Hogle, J. G. (1996). *Considering games as cognitive tools: In search of effective "edutainment."* Retrieved from http://twinpinefarm.com/pdfs/games.pdf

Klopfer, E., Osterweil, S., & Salen, K. (2009). *Moving learning games forward*. Cambridge, MA: The Education Arcade.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A. R., Evans, C., & Vitak, J. (2008). *Teens' gaming experiences are diverse and include significant social interaction and civic engagement*. Retrieved from http://www.pewinternet.org/Reports/2008/Teens-Video-Games-and-Civics.aspx

Mayrath, M. C., Clarke-Midura, J., Robinson, D. H., & Schraw (Eds.). (2012). *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishers.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Partnership for 21st Century Skills (2008). 21st century skills, education & competitiveness: A resource and policy guide. Retrieved from http://www.p21.org/storage/documents/21st_century_skills_education_and_competitiveness_guide.pdf

Pellegrini, A. D. (Ed.). (1995). *The future of play theory: A multidisciplinary inquiry into the contributions of Brian Sutton-Smith*. Albany, NY: State University of New York Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research & Development, 44*(2), 43–58.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Shute, V. J., & Becker, B. J. (2010). Prelude, Issues and assessment for the 21st century. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs* (pp. 1–11). New York: Springer-Verlag.

Shute, V. J., Gawlick, L. A., & Gluck, K. A. (1998). Effects of practice and learner control on short- and long-term gain and efficiency. *Human Factors, 40*(2), 296–310.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M-W. (in press). Advances in the science of assessment. To appear in *Educational Assessment*.

Shute, V. J., Rieber, L., & Van Eck, R. (2011). Games ... and ... learning. In R. Reiser & J. Dempsey (Eds.), *Trends and issues in instruction design and technology* (3rd ed., pp. 321–332). Upper Saddle River, NJ: Pearson Education.

IAP PROOF © 2017

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

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*, 423–430. doi:10.1080/00220671.2013.832970

----------> AU: Update?

Shute, V. J., & Wang, L. (in press). Assessing and supporting hard-to-measure constructs. To appear in A. Rupp, & J. Leighton (Eds.), *Handbook of cognition and assessment*.

Squire, K. D. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19–29.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied, 2*(4), i–109. doi:10.1037/h0092987

Ventura, M., & Shute, V. J. (2013). The validity of a game-based assessment of persistence. *Computers and Human Behavior*, *29*, 2568–2572.

Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky.* New York, NY: Plenum.

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L., ... Conkey, C. (2009). Relationships between game attributes and learning outcomes. *Simulation & Gaming*, *40*(2), 217–266. doi:10.1177/1046878108321866

IAP PROOF
© 2017

IAP PROOF
© 2017