Computers & Education 87 (2015) 340–356

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

# The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment

Yoon Jeon Kim [a, *], Valerie J. Shute [b]

[a] Worcester Polytechnic Institute, USA
[b] Florida State University, USA

ARTICLE INFO

*Article history:*
Received 26 April 2015
Received in revised form 6 July 2015
Accepted 13 July 2015
Available online 17 July 2015

*Keywords:*
Game-based assessment
Game-based learning
Psychometric evaluation
Validity
Reliability
Fairness
Gameplay sequences

ABSTRACT

Educators today are increasingly interested in using game-based assessment to assess and support students' learning. In the present study, we investigated how changing a game design element, *linearity in gameplay sequences*, influenced the effectiveness of game-based assessment in terms of validity, reliability, fairness, learning, and enjoyment. Two versions of a computer game, Physics Playground (formerly Newton's Playground), with different degrees of linearity in gameplay sequences were compared. Investigation of the assessment qualities—validity, reliability, and fairness—suggested that changing one game element (e.g., linearity) could significantly influence how players interacted with the game, thus changing the evidentiary structure of in-game measures. Although there was no significant group difference in terms of learning between the two conditions, participants who played the nonlinear version of the game showed significant improvement on qualitative physics understanding measured by the pre- and posttests while the participants in the linear condition did not. There was also no significant group difference in terms of enjoyment. Implications of the findings for future researchers and game-based assessment designers are discussed.



## 1. Introduction

The Committee on the Foundations of Assessment (Pellegrino, Chudowsky, & Glaser, 2001) asserted that advances in the cognitive and measurement sciences provide an opportunity to rethink current approaches to educational assessment. The Committee also suggested that new kinds of assessment should include authentic tasks that elicit evidence for what students know or can do, based on modern theories of learning, to provide teachers and students with actionable information. Similarly, Shute and Becker (2010) have called upon the education community for innovations in educational assessment by incorporating complex contexts where learning takes place alongside the assessment issues of validity, reliability, and fairness.

Responding to the need for new kinds of assessment, scholars and practitioners have been recognizing the potential of video games for assessment and learning (e.g., Baker & Delacruz, 2008; Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Using video games for assessment, however, poses challenges to educators and assessment designers (Mislevy et al., 2012; Zapata-Rivera & Bauer, 2012). Because both game and assessment design activities aim to create situations with which people interact, the two design processes can be compatible. Yet the two activities aim to

* Corresponding author.
*E-mail addresses:* yjkim@wpi.edu (Y.J. Kim), vshute@fsu.edu (V.J. Shute).

http://dx.doi.org/10.1016/j.compedu.2015.07.009
0360-1315/

achieve rather conflicting goals. That is, in games, the situations must be at the leading edge of what they can do to provide engagement and enjoyment, whereas in assessment, students' responses to these situations must provide evidence of their general proficiencies. Because game-based assessment (GBA) must address all of these considerations simultaneously, heavier emphasis on either one might hinder the enjoyment of the game or the strength of the evidence collected for assessment. Therefore, a well-designed GBA should be enjoyable to the players while maintaining the qualities of a good assessment.

Because the use of games as assessment is fairly new in education, scant evidence exists regarding how to balance design considerations of games vs. assessment to maximize the effectiveness of GBA without losing game-like characteristics such as fun and engagement. Furthermore, no empirical investigation has been conducted to evaluate the effectiveness of different game design choices in terms of assessment and learning. Therefore, an important question that the field needs to address is how to optimally balance assessment and game elements to (a) meet psychometric requirements of *validity*, *reliability*, and *fairness*, (b) improve learning, and (c) make it engaging for learners. One way to address this question is by conducting a series of A/B tests which compare different configurations of one game element controlling for other elements (Mislevy et al., 2014). A/B testing, commonly used in various design domains, is a form of experimental study in which two (or more) groups play games that are the same except for one particular element, then examine related metrics (e.g., psychometric qualities, enjoyment) to optimize design. Therefore, the purpose of this study is to investigate how specific design choices influence the effectiveness of GBA. Specifically, this study compares two versions of Physics Playground (formerly known as Newton's Playground) that differ in the linearity of gameplay sequences, and explores how manipulation of this element in the game may yield different results.

In the following sections, we first introduce our previous work with Physics Playground to describe how it is a valid tool to assess and support qualitative physics understanding of Newton's three laws. After that, we review the notion of linearity within the game literature and describe how we conceptualize linearity in gameplay sequences in the current study.

## 2. Previous work

Physics Playground (PP) is a computer-based game designed to simultaneously *assess* students' nonverbal understanding of physics principles, commonly referred to as qualitative physics. In PP, players draw various objects on the screen using a mouse, and once drawn, these objects become “alive” and interact with other objects. PP is characterized by an implicit representation of Newton's three laws of motion as well as the concepts of balance, mass, gravity and conservation of energy and momentum (Shute, Ventura, & Kim, 2013). These physics principles are operationalized by the use of simple mechanical devices, called *agents of force and motion*, such as levers, pendulums, and springboards. The primary mechanic of the game is to move a green ball to the red balloon on the screen by drawing and applying the four agents of force and motion. Players who solve a level earn either a gold or silver badge. A gold badge indicates an “elegant” solution (i.e., one with a limited number of objects, in most cases less than three objects). Therefore, acquiring gold badges is a proxy for mastery in the game. A silver badge simply means that the player has solved the problem, using more than three objects.

For example, Fig. 1 shows the level called “Trunk Slide.” When a player starts this level, the green ball slides along the elephant's trunk. To launch the ball up to the red balloon, the player may draw a springboard like the one shown in the right half of the figure. When a weight attached to the springboard (i.e., the green object) is released, the green ball flies up to hit the balloon.

The underlying assessment engine, built directly into the game, captures and analyzes the log data produced by the player's interactions with the game. This information is then used to make inferences regarding the player's understanding of qualitative physics, and to identify her strengths and weaknesses relative to various aspects of physics (See Shute, Ventura, & Kim, 2013 for a more detailed description of the game's design and assessments).

In a field study with 167 middle school participants (grades 8–9), we investigated whether (a) stealth assessment in PP satisfied construct validity, (b) playing PP could improve players' understanding of qualitative physics, and (c) players enjoyed the game (Shute, Ventura, & Kim, 2013). Regarding the first question, several in-game performance measures of PP (e.g., the

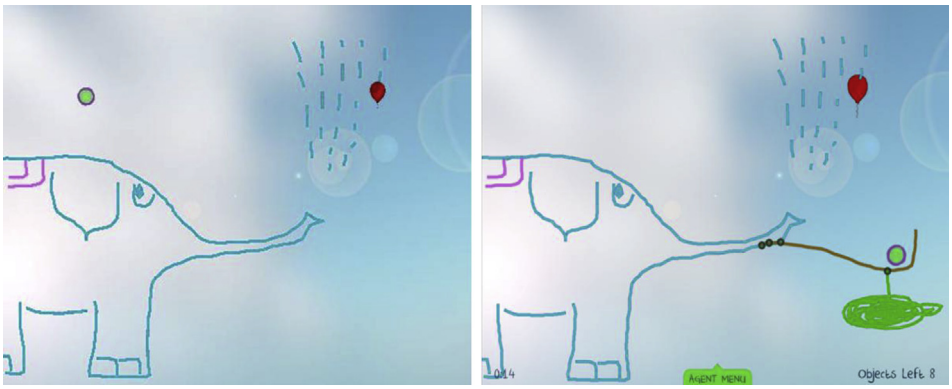


**Fig. 1.** Trunk Slide (on the left) and a springboard solution for the level (on the right).

total number of gold badges obtained across all levels) significantly correlated with the pretest and posttest of qualitative physics, suggesting construct validity of PP. Furthermore, gold badges yielded a stronger weight of evidence for physics proficiency than silver badge solutions. We also found a significant difference between the pretest and posttest scores, $t(154) = 2.12$, $p < .05$, as a function of playing PP. This finding suggested that some learning did occur as a function of gameplay. Regarding enjoyment, the participants enjoyed playing PP with the overall mean score of 3.80 ($SD = 1.10$) on a 5-point scale (with 1 being "strongly disagree" and 5 being "strongly agree").

In conclusion, these findings suggest that PP is a valid tool for assessing learning in a game. However, our findings were based on one set of game design decisions (e.g., a nonlinear gameplay sequence where players could move freely among different levels) because currently the principles that guide the "interplay of games with assessment" are lacking (Mislevy et al., 2012, p. 77). Therefore, the logical next step is a comparison of variations of the game to investigate how PP's different features and levels function in terms of assessment, learning, and enjoyment across different learners. These comparisons can ultimately lead to the development of a set of game-based assessment design principles. In the following section, we conceptualize one game design element that we investigated in the current study—*linearity in gameplay sequences*.

## 3. Linearity in the game literature

In general, linearity refers to the degree to which a game restricts player freedom or control (Rouse, 2005; Warren, 2009). Rouse (2005) identifies several dimensions of games in which game designers must determine the degree of linearity including (a) problem-solving actions and strategies within a level, and (b) gameplay sequences across levels. Although a low degree of linearity, broadly defined, is often associated with certain genres (e.g., role-playing and sandbox games), many game designers consider linearity and nonlinearity as opposite ends of a continuum in relation to interactivity and challenge (Rouse, 2005). Nonlinearity usually relates to high levels of enjoyment (e.g., Klimmt, Blake, Hefner, Vorderer, & Roth, 2009; Vorderer, Hartmann, & Klimmt, 2003). Similarly, Chen (2007) discusses the importance of allowing players to choose their own paths and levels of challenge in gameplay. Because players with different gaming abilites can experience different levels of difficulty at different times during gameplay, the availability of multiple gameplay paths based on ability is key to ensuring high levels of enjoyment for all levels of players.

Linearity, however, does not necessarily mean less enjoyment than nonlinear play. In fact, many games, such as *Portal 2* and *Uncharted*, are linear in terms of game sequences, yet highly engaging and enjoyable for many players. In addition, within the context of game-based assessment, linearity might provide several advantages over nonlinearity. First, linearity in game sequences across levels "forces" players to follow predetermined gameplay sequences ensuring content validity (i.e., covering a sufficient pool of levels that represent the construct). In contrast, when players choose different gameplay sequences in nonlinear games, these different paths essentially become different assessments (Almond, Kim, Velasquez, & Shute, 2014). Second, linearity in gameplay sequences might increase reliability of the assessment, especially when the game is not adaptive. That is, having different gameplay sequences among players can increase variance or decrease the "length" of the assessment, which can decrease reliability.

## 4. Psychometric qualities for game-based assessment

### 4.1. Validity

Validity in educational assessment broadly refers to the degree to which evidence and theory support the interpretation and uses of assessment, and validity is the most fundamental consideration in developing and evaluating assessment systems (AERA, APA, & NCME, 1999). While there are several sources of evidence for assessment validation (e.g., face validity) that can be considered in practice (Abell, Springer, & Kamata, 2009), the foremost source of validity, even considered as a unifying concept by Messick (1995), is construct validity. Construct validity refers to the extent to which a new assessment correlates with other assessments that measure the same construct (i.e., convergent construct validity) or does not correlate with dissimilar measures (i.e., divergent construct validity).

Although digital games often use rich, authentic problems that require players to apply their developing knowledge and skills, the problems still differ to varying degrees from real-world problems. To ensure validity in GBA, designers must be careful to include particular task features that are most crucial for eliciting evidence of the target knowledge and skills in the assessment situations (Mislevy et al., 2012). Moreover, the designed features of the game (e.g., players' interactions with the game, time spent on a specific task) need to be examined in relation to other measurements that capture the same construct.

### 4.2. Reliability

Reliability typically refers to the internal consistency with which test items measure overall proficiencies, and is typically calculated as various forms of intercorrelation coefficients (DeVellis, 2003; Mislevy et al., 2012). Reliability can also be viewed as the extent to which scores are free of measurement errors (AERA, APA, & NCME, 1999), which calls for assessment designers to identify and address possible errors in the early design and development stages (Abell et al., 2009).

### 4.3. Fairness

Fairness in educational assessment has four meanings: (a) lack of bias, (b) equitable treatment in the testing process, (c) equality in the use of outcomes from testing, and (d) equal opportunities for different subgroups to learn (AERA, APA, & NCME, 1999). Fairness should be addressed with particular care in GBA because such systems might function differently across subgroups (e.g., male vs. female, experienced gamers vs. non-gamers). For example, the literature generally reports that compared with females, males play all genres of games more frequently and for a longer duration, and are also more willing to sacrifice other activities to play games (Griffiths, 1997; Hamlen, 2010; Lucas & Sherry, 2004; Rideout, Foehr, & Roberts, 2010).

In addition, playing digital games leads to the development of other important skills such as computer literacy (Cassell & Jenkins, 1998), spatial abilities (Subrahmanyam & Greenfield, 1994), and cognitive and attentional skills (Green & Bavelier, 2012). Therefore, GBA should minimize the influence of the player's gaming ability relative to the accuracy of measuring his proficiency in the target skills and knowledge, and should ensure that different profiles of students are afforded equal opportunities to learn from GBA.

## 5. The current study

### 5.1. Research questions

In the current study, we examined the influences of linearity in gameplay sequences, on assessment quality, learning, and enjoyment. Our specific research questions included the following:

1. *Assessment*: Are the two versions of PP comparable in terms of validity, reliability, and fairness?
2. *Learning*: Are the two versions of PP comparable in supporting learning of qualitative physics? If not, which version is more effective?
3. *Enjoyment*: Are the two versions of PP comparable in terms of enjoyment? If not, which version is more enjoyable?

Addressing the first research question, we hypothesized that PP Linear would have better psychometric features than PP Nonlinear in terms of validity and reliability. The underlying assumption of this hypothesis was that PP Linear would provide more advantages in terms of these two qualities compared to PP Nonlinear, because the constraint to follow a predetermined gameplay sequence would minimize possible variations in gameplay patterns that might hinder assessment validity and reliability. However, we took an exploratory approach with investigation of fairness because we did not have strong theoretical support to hypothesize which version of PP would have more advantages.

Addressing the second research question, we hypothesized that the PP Linear group would show greater improvement in physics learning than the PP Nonlinear group. This was because, in the linear version, participants would be required to play equivalent numbers of levels for each agent of motion by following the predetermined sequence, whereas there was no guarantee that participants would do the same in the nonlinear version because they could freely choose which levels to play. Therefore, we further hypothesized that the PP Linear group would outperform the PP Nonlinear group on the qualitative physics posttest.

Addressing the third research question, we hypothesized that the PP Nonlinear group would enjoy playing the game more than the PP Linear group. This hypothesis was supported by the extant literature, which suggests that players prefer games that offer more freedom of choice and action, compared with games that offer constrained choices (e.g., Klimmt, Hartmann, & Frey, 2007; Ryan, Rigby, & Przybylski, 2006). Therefore, we expected players to enjoy NP Nonlinear more than NP Linear because they would be able to choose their own levels of play based on their own judgment of how fun or difficult a level appeared.

### 5.2. Participants and research design

We recruited 47 male and 55 female undergraduate students with an average age of 20.4 ($SD = 1.4$) who were enrolled at a large southeastern U.S. public university. While our prior work (e.g., Shute, Ventura, & Kim, 2013) recruited middle school students, we believe that the participants of the current study appropriately represent a population who needs educational support. That is, the literature shows that college students often have difficulty conceptually understanding Newton's laws, even after taking college-level Physics courses (e.g., Thornton & Sokoloff, 1998). In terms of gaming experience, as only four participants (3.9%) responded that they had played a game having similar mechanics to PP (e.g., *Crayon Physics Deluxe*, *Magic Pen*), the participants of the current study appropriately represent the broader population.

This study was a randomized design comparing two levels of an independent variable within Physics Playground (PP). The two levels were the two versions of PP with different degrees of linearity in gameplay sequences, as described above: (a) *PP Linear*, which constrained the sequence of gameplay across levels, and (b) *PP Nonlinear*, which did not have this constraint. A total of 52 and 50 participants were randomly assigned to either the PP Linear or PP Nonlinear condition, respectively.

### 5.3. Interventions

As noted, the two versions of the game provide different degrees of linearity in gameplay sequences. PP Linear has a total of five playgrounds with 4–6 levels per playground. Players assigned to this condition must follow a predetermined sequence

of play based on the difficulty of levels which was (a) estimated by difficulty indices formulated by the PP design team, and (b) prior empirical research on the overall difficulty of the agents (Shute, Ventura, & Kim, 2013). Table 1 displays all levels that were included in this study and their difficulty (p-values).

Within a playground, a player who earns a silver badge can choose to repeat the level if she wishes to try another way to solve it earning a gold badge, but she cannot skip to a harder level without first solving the easier levels. To move from one playground to the next, players must solve the required number of levels (e.g., 4 out of 6) in the easier playground. Fig. 2 shows the main menu screens of PP Linear and PP Nonlinear, and illustrates how the two versions have different interfaces for controlling gameplay sequences.

To maximize the comparability of PP Linear and PP Nonlinear in this study, both versions of the game included the same set of levels in the same order, and both versions allowed players to repeat any previously solved levels. However, PP Linear provided several smaller playgrounds compared to PP Nonlinear. That is, players were required to go through the playgrounds in a fixed sequence and solve a specified minimum number of levels to move to the next playground. In contrast, PP Nonlinear presented all the levels in one large playground, and players could move freely among levels as they wished. Thus, because players did not need to complete one level before moving to the next, it was unlikely that all players would end up following the same gameplay sequence. Each participant played their assigned version of the game for 2.5 h.

### 5.4. Measures

#### 5.4.1. Demographic survey

This demographic survey collected background information on the participants that may affect their performance in the game, thus affecting psychometric features, learning, and enjoyment. The survey included questions about participants' gender, GPA, and overall use of video games. See Appendix A for the full survey.

#### 5.4.2. Qualitative physics test

This test was designed to measure participants' understanding of qualitative physics. Working with physics experts, we originally developed a qualitative physics test (12-item test with matching forms, used in Shute, Ventura, & Kim, 2013). But in the current study, we used a revised version of this test consisting of 16 multiple-choice items within each of two equivalent forms (i.e., Form A and Form B). We conducted a reliability analysis on the equivalence of Form A and Form B. The Pearson's correlation coefficient between Form A and Form B was .76, and the Cronbach's alpha values for Forms A and B were .71 and .50, respectively. The test has four subscales that correspond to the four agents in PP (i.e., ramp, lever, pendulum, and springboard), and these subscales measure implicit understanding of underlying physics principles related to Newton's laws of motion, balance, mass, conservation of energy (potential and kinetic energy) and momentum, (Masson, Bub, & Lalonde, 2011; Riener, Proffitt, & Salthouse, 2005; Shute, Ventura, & Kim, 2013). All items require students to use qualitative

**Table 1**
Level sequence and difficulty.

| Sequence | Difficulty (p) | Agent | Name of level |
|---|---|---|---|
| 1 | 1.00 | Ramp | Lead the ball |
| 2 | 0.98 | Lever | Scale |
| 3 | 1.00 | Pendulum | On the upswing |
| 4 | 1.00 | Springboard | Yippie! |
| 5 | 1.00 | Ramp | Feed me |
| 6 | 0.98 | Lever | Not too far |
| 7 | 0.97 | Pendulum | Cloudy day |
| 8 | 0.98 | Springboard | Sunny day |
| 9 | 0.97 | Ramp | Swamp people |
| 10 | 0.96 | Lever | Need fulcrum |
| 11 | 0.95 | Pendulum | Tetherball |
| 12 | 0.89 | Ramp | Ultimate pinball |
| 13 | 0.77 | Lever | Work it up |
| 14 | 0.93 | Pendulum | Smiley |
| 15 | 0.86 | Springboard | Rollercoaster |
| 16 | 0.89 | Springboard | Dolphin show |
| 17 | 0.76 | Pendulum | Starry night |
| 18 | 0.74 | Springboard | Leprechaun |
| 19 | 0.76 | Ramp | Watch out |
| 20 | 0.72 | Lever | Up in the air |
| 21 | 0.61 | Ramp | Goal |
| 22 | 0.53 | Lever | Caveman |
| 23 | 0.64 | Pendulum | Freefall |
| 24 | 0.58 | Lever | Caveman |
| 25 | 0.45 | Pendulum | Can opener |
| 26 | 0.34 | Springboard | Double hoppy |
| 27 | 0.41 | Pendulum | Mr. Green |
| 28 | 0.19 | Springboard | Space travel |

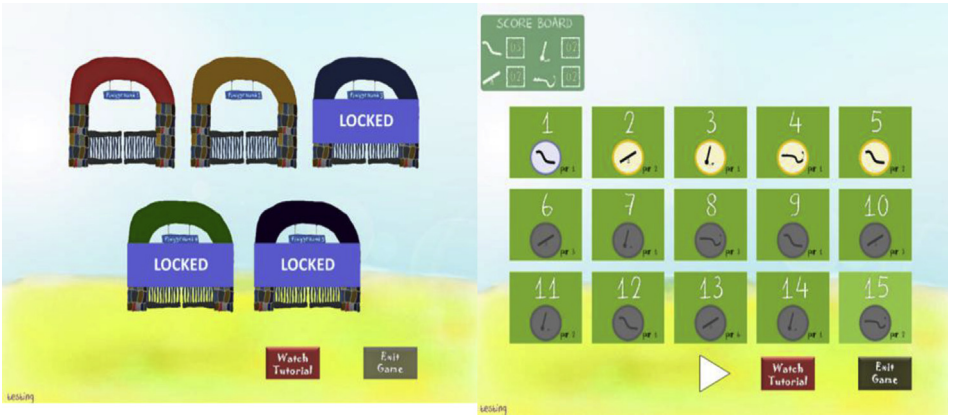


**Fig. 2.** PP Linear (left) and PP nonlinear (right).

reasoning about physical objects in their solutions, but do not focus on the explicit use of physics formulas or require formal physics knowledge. Fig. 3 is a sample item of the test.

*5.4.3. Game Play Questionnaire*

After participants finished playing the game and completed the physics posttest, they completed a modified version of the Game Play Questionnaire (GPQ). The GPQ is a subjective gameplay experience measure developed by Ryan et al. (2006). Based on the theoretical framework of self-regulated learning (SRL), Ryan et al. identified three factors that might influence players' enjoyment in video games: autonomy, competence, and intuitive controls. Autonomy refers to players' perceived degrees of freedom or number of choices in the game. Competence refers to players' feelings of being capable of solving problems within the game, and is related to the extent to which a game is challenging but not overwhelming. Ryan et al. (2006) reported that perceived autonomy and competence were associated with both game enjoyment and preferences for future play. Intuitive control refers to the ease by which players can learn how to play a game (i.e., the extent to which they can intuitively understand the game interface). In this study, we used a version of the GPQ that consisted of 15 Likert-scale items on a five-point scale (from 1 = Strongly disagree to 5 = Strongly agree). There are six items for enjoyment, three items for autonomy, three items for competence, and three items for intuitive controls. The survey showed reliability, Cronbach's alpha .88, that was similar to that of the original questionnaire and subscales (Ryan et al., 2006). See Appendix B for the full questionnaire.

### 5.5. Procedure

All participants played their assigned version of the game in a media lab located on the campus. There were 4–5 participants per each session, and each participant was provided with an individual laptop and headphone. While playing the game, participants rarely interacted with each other. All participants first completed the demographic survey and the qualitative physics pretest. To facilitate participants' familiarity with the game operation and mechanics, participants in both conditions (a) played an interactive tutorial to learn the mechanics of the game, (b) watched a short tutorial video about the four agents, and (c) played six warm-up levels that focused on the creation and use of the agents. The combined tutorials and warm-up session took up to 10 min to complete.

After the tutorials and warm-up session, the proctor read aloud instructions about the goal and rules of the game. These instructions informed the participants that their goal was to score as high as possible by solving as many levels as possible. They were also instructed that they can score higher by earning gold badges, which counted as 2 points (i.e., double the score of silver trophies). Although the goal was not reiterated during the session, participants could see their current score by checking the scoreboard located on the top of the screen.

Depending on their assigned treatment condition, participants then played either the linear or nonlinear version of PP for 2.5 h. Seventeen participants out of the total sample of 102 (9 for PP Linear and 8 for PP Nonlinear) finished all 28 levels within 2.5 h. To keep the gameplay time constant, however, those players who finished early were encouraged to continue playing, and those players went back to the earlier levels in which they only earned silver badges and replayed them to achieve higher scores via obtaining gold trophies. After the 2.5 h of gameplay, all participants completed the qualitative physics posttest and the GPQ. The full procedure took three hours, and was completed within one session.

## 6. Results

### 6.1. Validity

For the first research question, we investigated whether the two versions of the game were comparable in terms of assessment qualities, specifically validity. We hypothesized that PP Linear would have better assessment qualities related to

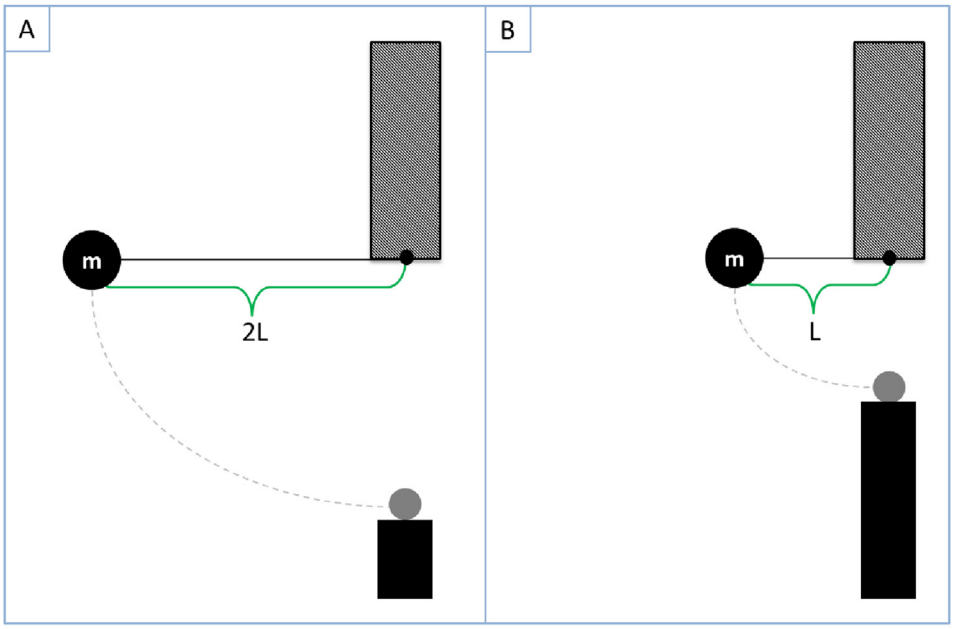


**Fig. 3.** A sample item of the qualitative physics test.

validity than PP Nonlinear. In a previous study, the Shute, Ventura, and Kim (2013) reported that in-game performance measures (e.g., the total number of gold badges across all levels) were significantly correlated to an external measure of physics understanding, suggesting construct validity, whereas silver badges were not (except for the silver badges for springboard solutions). The version of the game used in the 2013 study was similar to PP Nonlinear in terms of gameplay sequences since it allowed players to skip around as desired. For the present study, to investigate the validity of the in-game measures of PP for qualitative physics understanding, we calculated partial correlation coefficients between three in-game measures (i.e., completion rate, number of gold badges, and number of silver badges) and physics posttest scores for each condition, after controlling for incoming physics understanding.

As shown in Table 2, the completion rate—the total number of solved problems (either silver or gold badge solutions) divided by the total number of available levels—demonstrated similar strengths of evidence for both conditions. However, in the PP Linear condition, the total number of silver badges provided stronger evidence for physics understanding than the total number of gold badges. Conversely, in PP Nonlinear, only the total number of gold badges provided positive evidence for physics understanding, and silver badges were associated with slightly lower levels of physics understanding.

We then transformed the correlation coefficients to z-scores to evaluate whether any of the differences among the correlation coefficients were statistically significant. Only the correlation coefficient difference between silver badges for PP Linear and PP Nonlinear was significant ($z = 1.94$, $p = .05$) indicating the evidentiary strength of an earned silver badge for physics understanding was significantly different between the two versions of the game. We further calculated correlations between the number of agent-specific silver and gold badges and the physics posttest scores, and then transformed the correlation coefficient differences to z-scores (Table 3).

In PP Nonlinear, the number of springboard badges was the only in-game measure that was significantly associated with physics understanding for both silver ($r = .27$) and gold ($r = .48$) badge levels, while other agent-specific silver badges did not provide any evidence for physics understanding. Similarly, the Shute, Ventura, & Kim, 2013 reported that the number of springboard-specific badges was associated with high levels of qualitative physics understanding (i.e., $r = .29$ for silver badges and $r = .49$ for gold badges, $p < .05$). In PP Linear, however, all agent-specific badges, both silver and gold, were associated with physics understanding except for the gold springboard (which was marginally significant at $p = .08$). Agent-specific silver badges generally displayed stronger associations with physics understanding than gold badges in the PP Linear version of the game.

### 6.2. Reliability

We initially hypothesized that PP Linear would demonstrate better reliability than PP Nonlinear. Using eight agent-specific in-game observables (i.e., number of silver and gold badges for ramp, lever, pendulum, and springboard), we first calculated

**Table 2**
Correlations between posttest scores and physics playground badges controlling for pretest scores.

| | Completion rate | Silver badges | Gold badges |
|---|---|---|---|
| PP linear ($n = 49$) | 0.37** | 0.29* | 0.20 |
| PP nonlinear ($n = 47$) | 0.39** | −0.11 | 0.42** |
| z-score | −0.11 | 1.94* | −1.16 |

**$p < .01$, *$p < .05$.

the Cronbach's alpha coefficient for each version of the game. According to some researchers (e.g., Raykov & Shrout, 2002; Yang & Green, 2011), Cronbach's alpha is a sub-optimal measure of reliability for such data because (a) it is not a good measure of internal consistency, and (b) it underestimates reliability when the underlying latent construct is not unidimensional. Nevertheless, we calculated alpha coefficients because they provided useful baseline values. The coefficients were calculated from the eight in-game observables for PP Linear and PP Nonlinear, yielding values of .63 and .50, respectively.

To further evaluate the reliability of in-game measures of both versions PP, particularly agent-specific silver and gold badges to provide evidence for players' physics understanding, we followed a two-step structural equation model (SEM) approach as described in Yang and Green (2011). The first step in this procedure required evaluation of several confirmatory factor analytic (CFA) models to establish the most appropriate model that (a) was consistent with our understanding of the construct of interest, and (b) provided an acceptable model-data fit.

Several types of fit indices are commonly used to judge the goodness of model-data fits: comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). According to Hu and Bentler (1999), CFI values range between 0 and 1, and any value greater than .95 indicates a good data-model fit even if the sample size is small. When a model fit is perfect, the RMSEA value becomes 0. Therefore, any RMSEA value smaller than .06 indicates a good data-model fit, and any value larger than 0.10 might indicate a poor fit (Browne & Cudeck, 1992). However, Hu and Bentler (1999) also caution that RMSEA values can be overestimated when the sample size is small. SRMR values range between 0 and 1, and any value smaller than .08 indicates a good data-model fit.

Using Mplus Version 6.11 (Muthén & Muthén, 2007), we first considered a one-factor model (i.e., silver and gold badges combined as the top-level factor) with the variance of the factor fixed at 1. For both PP Linear and PP Nonlinear, this model fit the data poorly. Specifically, for PP Linear, $\chi^2(20) = 227.73$, $p < .001$, CFI = .35, RMSEA = .45, and SRMR = .20; for PP Nonlinear, $\chi^2(20) = 145.04$, $p < .001$, CFI = .67, RMSEA = .35, and SRMR = .18. Next, we evaluated a two-factor model (i.e., silver and gold badges as the top-level factors) with the variance of the factors fixed at 1, and allowance for the two factors to be correlated. Although this model fit the data slightly better than the one-factor model, the fit was still poor. For PP Linear, $\chi^2(19) = 166.01$, $p < .001$, CFI = .54, RMSEA = .39, and SRMR = .10; for PP Nonlinear, $\chi^2(19) = 74.04$, $p < .001$, CFI = .85, RMSEA = .24, and SRMR = .10.

Based on the modification indices suggested in the output of the second model (i.e., two-factor model), we added covariance between errors to Model 2 by allowing errors to be correlated. Fig. 4 shows our final model for PP Linear and PP Nonlinear.

We used our final model to evaluate reliability of the game by calculating ω reliability coefficients because (a) it shows a reasonable fit for both PP Linear and PP Nonlinear as indicated in Table 4 (i.e., CFI 0.95 and 0.96 for PP Linear and PP Nonlinear, respectively), (b) the structure of the model is aligned with our understanding of how players pursue silver and gold badges differently, and (c) this model meets the configural invariance requirement (i.e., the same model fits across the two groups). We should note that we primarily used CFI to evaluate data-model fit in this study because other chi-square based indices (e.g., RMSEA) can be highly biased when a sample size is small (Kenny, Kaniskan, & McCoach, 2014). Similar to the Cronbach's alpha values, the ω value for PP Linear was higher than that of PP Nonlinear, .96 [95% CI = .94, .97] and .92 [95% CI = .89, .95], respectively. Although both reliability values are fairly high, ω for PP Linear has a shorter span for the 95% Confidence Interval compared to PP Nonlinear indicating, overall, the reliability of PP Linear is slightly better than PP Nonlinear (Cheung, 2009).

To further test why the reliabilities between the two versions of PP were not as different as expected, we analyzed gameplay sequences using R (R Core Team, 2013) and an R-package called TraMineR (Gabadinho, Ritschard, Müller, & Studer, 2011). For this analysis, we created a sequence plot for each version of the game to visually inspect how gameplay sequences differed across the two versions (Fig. 5). Each color block in the plots represents a unique level in the game, and the colors follow a gradient from blue to red. That is, the darkest blue is the first level in the first playground and the darkest red is the last level in the last playground. As shown in Fig. 5, participants in both groups started with easier levels and gradually moved to harder levels. Therefore, the patterns did not differ much between PP Linear and PP Nonlinear—that is, until players reached later levels that were much more difficult. This “jumping around” became more apparent in the later levels of PP Nonlinear.

**Table 3**
Correlations between posttest scores and agent-specific badges.

| | Silver RA | Silver LE | Silver PE | Silver SB | Gold RA | Gold LE | Gold PE | Gold SB |
|---|---|---|---|---|---|---|---|---|
| PP linear ($n = 52$) | 0.32* | 0.50** | 0.34* | 0.45* | 0.29* | 0.27* | 0.38** | 0.24 |
| PP nonlinear ($n = 50$) | 0.06 | 0.11 | −0.03 | 0.27* | 0.41** | 0.44** | 0.55** | 0.48** |
| z-score | 1.33 | 2.15* | 1.88 | 1.02 | −0.67 | −0.96 | −1.07 | −1.36 |

**$p < .01$, *$p < .05$, RA = Ramp, LE = Lever, PE = Pendulum, SB = Springboard.

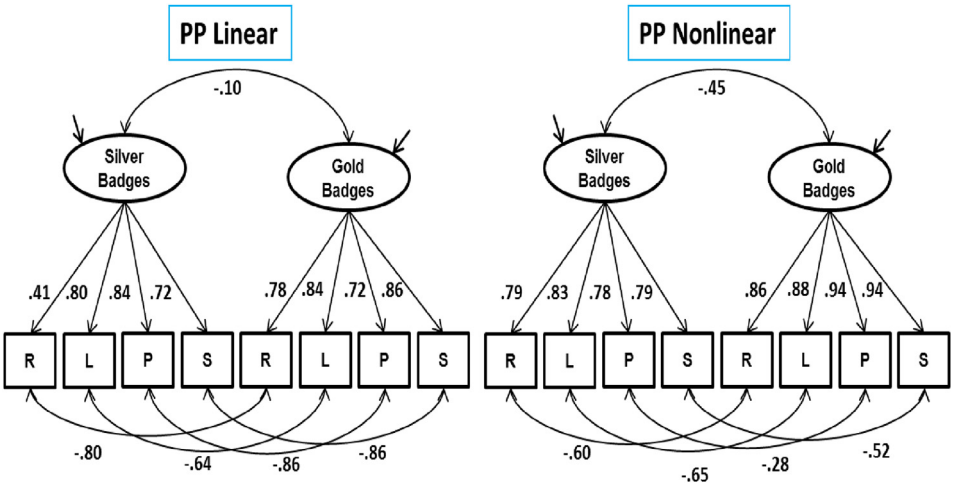


**Fig. 4.** Structural model for PP Linear (left) and PP Nonlinear (right) with configural invariance.

We then calculated normalized longitudinal entropy values for both conditions to compare how often this "jumping around" occurred in PP Nonlinear. A normalized longitudinal entropy value, which ranges from 0 to 1, can be calculated by summing the number of transitions between states (i.e., levels in this case) and then normalizing the value by dividing it by the maximum possible number of transitions. Thus, an entropy value closer to 0 indicates fewer transitions between levels and more repetitions of a same level. Table 5 summarizes descriptive statistics for the entropy values for each condition. The means of the entropy for PP Linear and PP Nonlinear were quite similar: .92 ($SD = .08$) for PP Linear and .93 ($SD = .05$) for PP Nonlinear.

This finding supports our interpretation from visual inspection of the sequence plots that players in PP Nonlinear did not greatly deviate from the predetermined sequence until they reached more difficult levels later in the game. We should note, however, that the entropy values for PP Nonlinear were more densely distributed on the right tail of the distribution than PP Linear (refer to the histograms in Fig. 6).

### 6.3. Fairness

Our next focus examined whether the two versions of the game were comparable in terms of fairness, particularly in relation to gamers vs. non-gamers and males vs. females. As we did not have strong theoretical or empirical evidence to hypothesize which condition would be better in terms of fairness, we took an exploratory approach. In the demographic survey, we included three questions that could help to distinguish gamers from non-gamers. The first question asked about game play *frequency* on a 7-point Likert-type scale (i.e., 1 = Not at all, to 7 = Every day for more than 3 h). The second question asked about *preferences* for playing video games over other activities on a 5-point Likert-type scale (i.e., 1 = Never, 5 = Often). The third question asked about *self-perceptions* related to playing video games on a 5-point Likert-type scale (i.e., 1 = non-player, 5 = expert). All three questions were significantly correlated to each other and to the number of gold badges with the two conditions combined (see Table 6).

Because gameplay frequency was highly correlated with self-perception as a gamer and the number of gold badges achieved in the game, we chose the game frequency question to create a categorical variable with two categories for gamers vs. non-gamers (labeled "gaming background" in subsequent analyses). That is, around 50% of the participants who fell into the range between 1 and 3, inclusive (1 = Not at all, 2 = About once a month, 3 = A few times a month) were coded as non-gamers, and the rest of the participants who fell into the higher ranges (4 = A few times a week, 5 = Every day, but less than 1 h, 6 = Every day, for 1–3 h, 7 = Every day, for more than 3 h) were coded as gamers.

Table 7 displays correlations between the gaming background variable (i.e., non-gamer = 0 vs. gamers = 1) and three in-game measures (i.e., completion rate, the total number of silver badges, and the total number of gold badges) after controlling for the pretest. For both PP Linear and PP Nonlinear, correlations between the gaming background and gold badges variables were significant ($r = .36$, $p = .01$; $r = .52$, $p < .001$, respectively). This finding suggests that gamers may have had an unfair advantage over non-gamers for achieving gold badges, regardless of their incoming physics knowledge. Additionally, gamers in PP Nonlinear had a slightly greater advantage over gamers in PP Linear in relation to gold badges obtained, although the difference was not significant ($z = -.95$, $p = .34$). For silver badges, neither PP Linear nor PP Nonlinear showed an unfair advantage for gamers.

We also investigated the effects of gender on various in-game measures after controlling for physics understanding (Table 8). By simply looking at the completion rate, it might appear that PP Nonlinear is more biased against female players ($r = -.38$, $p < .001$) than PP Linear ($r = -.14$, $p = .34$). However, there were clear gender differences in PP Linear relative to the type of badges that female and male players worked towards. That is, female players (coded as "1") mainly obtained silver badges ($r = .28$, $p = .05$) while male players (coded as "0") appeared to focus on obtaining gold badges ($r = -.53$, $p < .001$).

**Table 4**
Fit Indices for Model 3 and ω Coefficients.

| | $\chi^2$ | df | CFI | RMSEA | SRMR | $\omega$ |
|---|---|---|---|---|---|---|
| PP linear | 31.42 | 15 | 0.95 | 0.15 (90% CI = 0.07, 0.22) | 0.097 | 0.96 (95% CI = 0.93, 0.98) |
| Nonlinear | 30.70 | 15 | 0.96 | 0.14 (90% CI = 0.07, 0.22) | 0.108 | 0.92 (95% CI = 0.89, 0.96) |

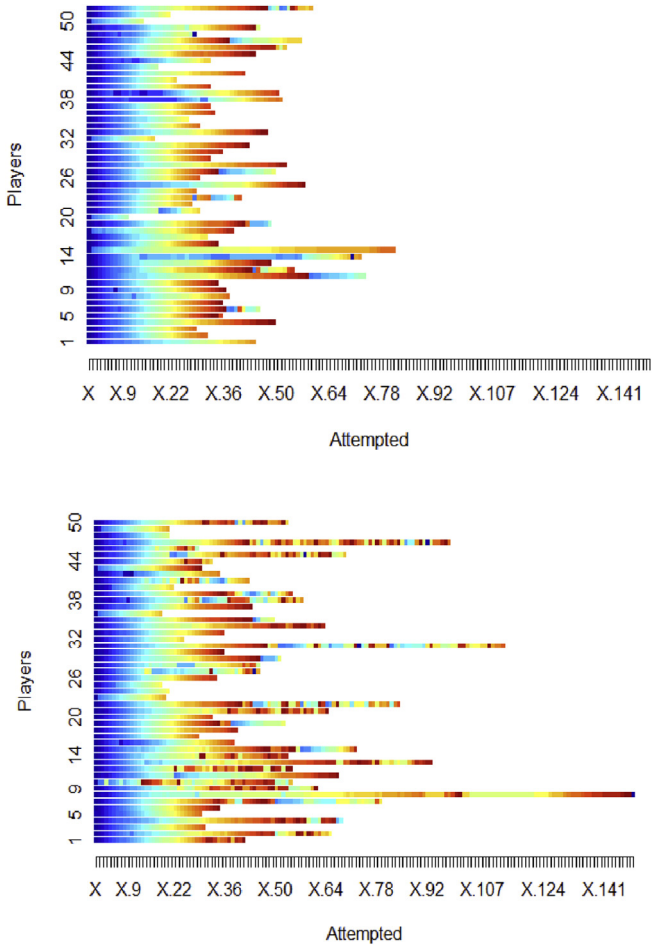


**Fig. 5.** Sequence plot for PP Linear (top) and PP Nonlinear (bottom).

Our next analysis aimed to determine whether the influences of the grouping variables (i.e., gender and gaming background) on the two focal in-game measures (i.e., the total number of gold and silver badges) differed between the two versions of PP controlling for players' incoming physics understanding (using physics pretest score as a covariate). We conducted four 2-way ANOVAs with pretest score as a covariate. The specifications for the four models and their results are summarized in Table 9.

Both Models 1 and 2, with gold badges as the outcome measure, yielded significant main effects for the two factors. The interaction for Model 1 (i.e., treatment condition with gender) was not significant, whereas the interaction for Model 2 (i.e., treatment condition with gaming background) was marginally significant ($p = .07$). Neither Model 3 nor Model 4 with silver badges as the outcome measure, yielded significant main or interaction effects. In summary, these findings suggest that for both the PP Linear and PP Nonlinear conditions, males who are also gamers may have an unfair advantage over females who are not gamers in terms of obtaining gold badges.

**Table 5**
Descriptive statistics of entropy for PP linear and PP nonlinear.

| | *M* (*SD*) | Min. | Max. |
|---|---|---|---|
| PP linear ($n = 52$) | 0.92 (0.08) | 0.67 | 1.00 |
| PP nonlinear ($n = 50$) | 0.93 (0.05) | 0.82 | 1.00 |

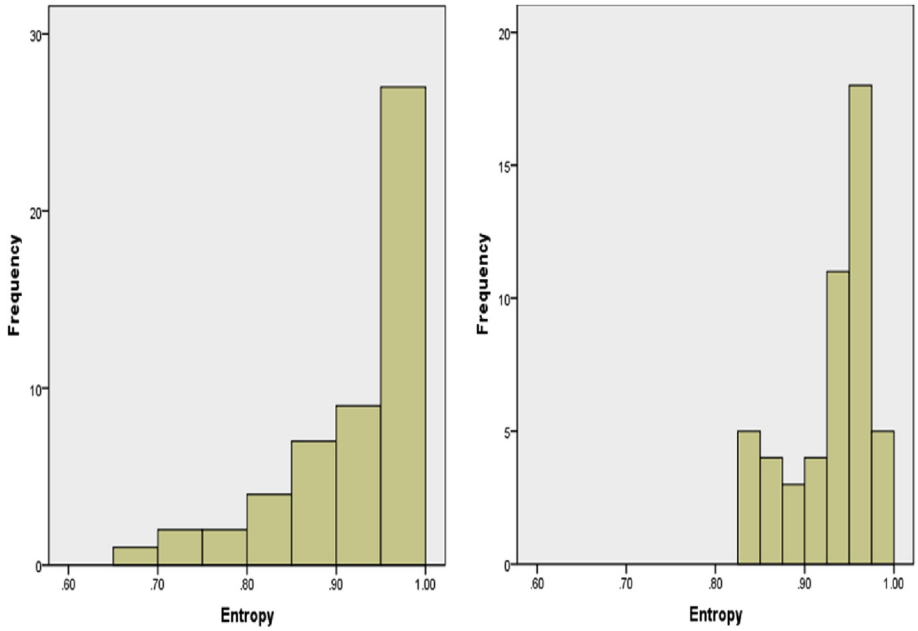


**Fig. 6.** Longitudinal entropy histogram for PP Linear (left) and PP Nonlinear (right).

### 6.4. Learning

For the second research question, we investigated whether the two versions of the game were comparable in terms of promoting learning of physics concepts. We initially hypothesized that PP Linear would be more effective regarding learning because players would be more likely to play a full range of problems following the fixed sequence compared to the PP Nonlinear players (who were free to move around to any levels they wished). That is, when players had full control over gameplay sequences, it was possible that they would skip problems that they found difficult without fully engaging with them, and this could unbalance the intended content of the game.

To minimize the possibility of order effects, we used two isomorphic forms of the physics test. One group took Form A as the pretest and Form B as the posttest while the other group received the tests in the opposite order. We first conducted a reliability analysis on the equivalence of Form A and Form B. The Pearson's correlation coefficient between Form A and Form B was .76, and the Cronbach's alpha values for Forms A and B were .71 and .50, respectively. Because pretest scores indicated that Form B was more difficult than Form A, we equated the scores using the linear linking function (Holland & Dorans, 2006; Holland, Dorans, & Petersen, 2007). The equated scores were used as the physics pre- and posttest scores in the subsequent analyses.

We first examined whether the two conditions had equivalent incoming physics understanding by calculating a t-test. Although the pretest scores were higher for the PP Linear condition ($M = 10.81$) than for the PP Nonlinear condition ($M = 9.93$), the difference was not significant ($t(100) = 1.88$, $p = .06$). Overall, when the two conditions were combined, there was a significant pretest-to-posttest gain with a medium effect size as a function of playing the game for 2.5 h ($t(101) = 3.03$, $p < .001$, $d = 0.60$). Contrary to our initial hypothesis, however, there was no significant mean score difference between the two versions of PP on the physics posttest, $t(100) = 0.82$, $p = .41$, $d = 0.16$. Furthermore, only the learning gain in the PP Nonlinear condition was significant with a large effect size ($t(49) = 2.95$, $p < .001$, $d = 0.84$).

Because the investigation of validity and fairness revealed relationships between physics understanding relative to the two grouping variables (i.e., gender and gaming background), we examined whether the difference between learning gains by condition was mediated by gender or gaming background. Therefore, we conducted two-way repeated measures ANOVAs adding gender and gaming background as between-subject factors. The interaction between the treatment condition and gender was significant ($F(3, 98) = 3.16$, $p = .03$, $\eta^2 = 0.09$) while the interaction between the treatment condition and gaming background was not significant ($F(3, 98) = 2.18$, $p = .10$, $\eta^2 = 0.06$). Fig. 7 illustrates the adjusted pretest and posttest scores for each condition and gender.

**Table 6**
Correlations among gaming-related questions.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Game frequency | | 0.40** | 0.73** | 0.48** |
| 2. Games vs. other activities | | | 0.54** | 0.20* |
| 3. Self-perception | | | | 0.45** |
| 4. Number of gold badges | | | | |

**$p < .01$, *$p < .05$.

**Table 7**
Partial Correlations between Gaming Background and in-Game Measures.

| | Completion rate | Silver badges | Gold badges |
|---|---|---|---|
| PP linear ($n = 49$) | 0.29* | 0.04 | 0.36** |
| PP nonlinear ($n = 47$) | 0.50** | −0.07 | 0.52** |
| z-score | −1.19 | 0.52 | −0.95 |

**$p < .01$ *$p < .05$.

**Table 8**
Correlations between gender and in-game measures controlling for pretest.

| | Completion rate | Silver badges | Gold badges |
|---|---|---|---|
| PP linear ($n = 49$) | −0.14 | 0.28* | −0.53** |
| PP nonlinear ($n = 47$) | −0.38** | −0.05 | −0.38* |
| z-score | 0.23 | 1.60 | −0.90 |

**$p < .01$ *$p < .05$; males = 0; females = 1.

### 6.5. Enjoyment

For the third research question, we tested if the two versions of PP were comparable in terms of enjoyment. We initially hypothesized that PP Nonlinear would be more enjoyable than PP Linear because it allowed more freedom and control regarding gameplay sequences. The literature supports the notion that perceived levels of freedom and control within a game are associated with enjoyment (Klimmt et al., 2009; Ryan et al., 2006; Vorderer et al., 2003). Since the revised GPQ included six items for enjoyment and three additional subscales (i.e., for autonomy, competence, and intuitive controls), we first computed a one-way MANOVA including all four composite scores as the outcomes. There was no statistically significant difference in the four aspects of the GPQ (Wilks' $\lambda = .96$, $F(1, 101) = 1.09$, $p = .30$, $\eta^2 = 0.04$), indicating that none of the measures in the questionnaire differed across the two conditions.

## 7. Discussion

As displayed in Table 2, the total number of silver badges a player attained provided stronger evidence for physics understanding than the number of gold badges in PP Linear. Conversely, only the number of gold badges a player attained provided evidence for physics understanding in PP Nonlinear. This finding suggests that changes in just one design element, in this case linearity in gameplay sequences, can significantly impact players' interactions with the game by changing players' mental "operational rules" during play. It further suggests that changing such a design element can ultimately influence the in-game indicators' weight of evidence in GBA, thus affecting validity. That is, although the articulated goal of the game was "to achieve the highest score possible," for both the PP Linear and PP Nonlinear conditions, the goal of PP Linear essentially became "to unlock the next level." Therefore, players in the PP Linear condition might have focused less on scoring high points by trying to get gold badge solutions, and more on unlocking additional levels. Players who play linear games, such as *Angry Birds* and *Candy Crush*, often display similar behaviors since they focus simply on solving levels to progress. In PP Nonlinear, players had complete control over their choices for the next level to play, so the operational rule of PP Nonlinear remained, "to get a high score." This goal was often achieved by selecting levels that could easily be solved with a gold badge solution.

We further investigated correlations between eight in-game measures (specific to the four agents) and players' scores on the physics posttest. As shown in Table 3, all eight in-game measures provided evidence for physics understanding in PP Linear (although the gold springboard badge data were marginally significant). Interestingly, for PP Nonlinear, the springboard silver badge data was the only agent-specific silver badge data that still provided evidence for physics understanding. This finding is consistent with what was reported in the previous study (Shute, Ventura, & Kim, 2013). One possible explanation for this is that springboard problems are more difficult to execute than problems for other types of agents. Thus, even earning a silver badge for a springboard problem may be sufficient to yield evidence for physics understanding.

Regarding reliability, although the overall reliability of PP Linear was slightly better than PP Nonlinear, the difference was smaller than we expected. One possible explanation for this finding is that players naturally followed the default order, since

**Table 9**
Four two-way ANOVA models used for fairness investigation.

| Model | Model specifications | Main | Interaction |
|---|---|---|---|
| 1 | Gold badges ~ condition + gender + pretest + condition*gender | $F(1, 101) = 6.03$, $p = .02$<br>$F(1, 101) = 21.86$, $p < .001$ | $F(1, 101) = 0.12$,<br>$p = .73$ |
| 2 | Gold badges ~ condition + gaming background + pretest + condition*gaming background | $F(1, 101) = 3.92$, $p = .05$<br>$F(1, 101) = 25.10$, $p < .001$ | $F(1, 101) = 3.45$,<br>$p = .07$ |
| 3 | Silver badges ~ condition + gender + pretest + condition*gender | $F(1, 101) = 0.15$, $p = .70$<br>$F(1, 101) = 0.47$, $p = .49$ | $F(1, 101) = 1.34$,<br>$p = .25$ |
| 4 | Silver badges ~ condition + gaming background + pretest + condition*gaming background | $F(1, 101) = 0.28$, $p = .60$<br>$F(1, 101) = 0.03$, $p = .86$ | $F(1, 101) = 0.48$,<br>$p = .50$ |

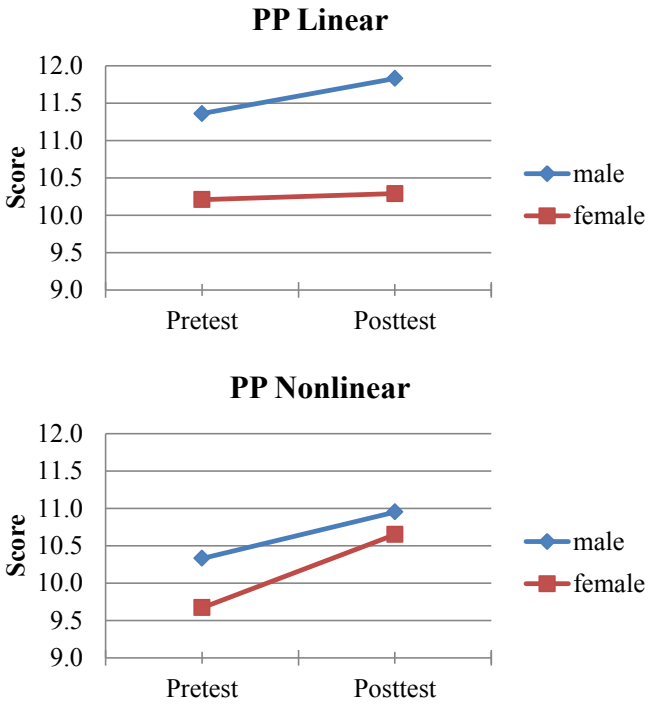


**Fig. 7.** Pretest and posttest scores for each condition by gender.

that is the typical method of playing commercial games, and only skipped levels when they needed to use the nonlinearity feature strategically to solve more problems. Another possible reason for the reliability finding is because the game used in this study included only 28 levels, which might not be long enough to see much variation in gameplay sequences. In conclusion, from the reliability and sequence analyses, linearity in gameplay sequences does appear to influence the reliability of game-based assessment. Moreover, the linear sequence design might be more beneficial in GBA, particularly if the primary goal is to achieve robustness of assessment, since it has slightly better reliability than the nonlinear sequence design.

The findings from the fairness investigation are somewhat perplexing. Although males had a clear advantage in terms of achieving gold badges in both conditions, females earned more silver badges than gold badges only in PP Linear. It appeared that there was a "gender divide" in the achievement of gold and silver badges in PP Linear. This gender divide in PP Linear could be explained by the extant research on how male and female gamers play games to satisfy different psychological needs and demonstrate different goal orientations. Broadly speaking, male players tend to have a higher need for challenge and achievement than female players, and show a stronger orientation towards extrinsic motivation than intrinsic motivation (e.g., Boyle & Connolly, 2008; Magerko, Heeter, & Medler, 2010). Applying this perspective, it might be possible that female players in PP Linear paid less attention to the scoreboard than male players, and made less effort to get gold badges. It is also possible that this gender divide was more apparent in PP Linear than in PP Nonlinear because the goal of PP Linear essentially became unlocking levels rather than achieving a high score.

Regarding learning, the t-test results revealed that only the participants in the PP Nonlinear condition showed a significant learning gain from pretest to posttest with a large effect size. However, the posttest difference between PP Linear and PP Nonlinear was not significant. A follow-up analysis using gender as another between-subject factor revealed a significant interaction between gender and treatment condition, indicating that females appeared to benefit from Nonlinear PP more than Linear PP (see Fig. 7). One possible explanation for this finding is related to the results revealed in the validity test. That is, the question arises about whether players' goals in PP Linear essentially became unlocking the next levels rather than striving for gold badges. This might have limited players' opportunities to explore different problems or solutions, and decreased their motivation to try to achieve efficient solutions aiming for gold badges.

Regarding enjoyment of the game by the players, the results from the MANOVA revealed no significant difference between the two conditions on the mean scores of enjoyment, autonomy, competence, or intuitive controls. There are several possible explanations for this finding. First, because this study was a controlled experimental study, participants were required to play the game for 2.5 h, even if they did not like the game, which might have affected players' overall enjoyment. Second, although the two game versions differed in terms of the degree of linearity in gameplay sequences, all other elements were identical (e.g., amount of control given within a level). It is possible that the players who played PP Nonlinear did not fully utilize the feature of nonlinearity feature (as discussed in the Reliability section of this paper), and did not experience more freedom and choices than the players assigned in PP Linear.

## 8. Implications

In the current study, we aimed to investigate the influence of one game element, linearity, on assessment qualities (i.e., validity, reliability, and fairness), learning, and enjoyment in game-based assessment (GBA). This study was motivated by the need to establish design principles specific for GBA, and the first step toward meeting that goal is to conduct a series of A/B testing experiments. These experiments involved the manipulation of one game element while keeping all other game elements the same, and the evaluation of how this manipulation influenced different aspects of GBA.

The validity investigation revealed that changes in players' degrees of control concerning gameplay sequences could change the operational goal of the game, and thus change the evidentiary structure and strength of the in-game measures. This finding has several implications for the field. First, GBA designers should conduct a series of play-testing studies, as early and as often as possible, to understand how different game design decisions ultimately change players' goals and behaviors within the game. Second, assessment designers must use the information obtained from play-testing to iteratively build assessment mechanics (e.g., scoring rules, evidence accumulation methods).

Contrary to the common belief that gamers prefer games with features offering more freedom and control, the reliability investigation revealed that players would not use those features unless it served a specific purpose (e.g., earning more gold badges). That is, as revealed in the sequence analysis, participants who played PP Nonlinear did not use the nonlinearity feature in easier levels because they could progress without using it. Thus, to design a game for the purposes of assessment and learning, the designer should carefully consider *how players would use the player-control feature and for what purposes, and how having that feature would influence assessment and learning.*

The findings from the fairness investigation also provide implications for the field concerning how designers should address the issue of measuring gaming ability vs. competencies of interest (Mislevy et al., 2014). Gamers and males may potentially have unfair advantages over non-gamers and females in the context of GBA, and the extent of these advantages, specific to certain in-game measures, may easily be influenced by game features. One implication for the field, which we presented in the learning analysis, is that if the primary goal of GBA is learning, rather than assessment, it might be acceptable to provide more control to the players, thus allowing them to explore as long as validity and reliability are established. In addition, it may be possible that learning occurs in the context of GBA when players try different strategies and solutions rather than achieving success quickly.

Finally, the investigation of enjoyment in both PP Linear and PP Nonlinear did not show any difference between the two versions of the game. This was somewhat surprising since the extant literature emphasizes the importance of providing players with sufficient control and freedom to facilitate enjoyment and engagement. We believe qualitative approaches (e.g., think-aloud protocol and interviews) combined with different metrics (e.g., the Game Play Questionnaire) could shed light on this issue.

## 9. Limitations

This study had several limitations related to its research design. First, we maintained a constant gameplay time–all participants played the game for 2.5 h within one session. However, when people play games for fun, they play for as long as they wish, in whatever intervals they wish. Therefore, understanding how people interact with a game in a controlled setting may not be the optimal approach to understanding in-game behaviors, particularly in terms of players' enjoyment. Second, we investigated only one type of linearity in this study, linearity in gameplay sequences, and the findings from this study do not provide insights for how other types of linearity (e.g., differential control permitted within a level) would influence assessment, learning, and enjoyment. Therefore, future work should involve different configurations of a game manipulating different types and levels of linearity to fully understand the issue of control (or choices) in the context of game-based assessment. Finally, the qualitative physics test, especially Form B, did not have a high degree of reliability. Therefore, future work should require refinement of the physics test.

## 10. Conclusion

In conclusion, we attempted to answer the question of how to balance game and assessment elements while simultaneously satisfying assessment, learning, and enjoyment requirements by investigating one game design choice, linearity in gameplay sequences, in the current study. The results were somewhat mixed, and it was also challenging to provide explanations for what might have happened during gameplay. Future studies should further explore how different game elements (e.g., challenge, reward systems) influence assessment, learning, and enjoyment. The results can then contribute to the establishment of design principles specific to game-based assessment.

## Acknowledgments

This research was funded by the Bill and Melinda Gates Foundation (#OPP1035331). Any opinions expressed are those of the authors and do not necessarily reflect the views of the funding agency.

## Appendix A. Demographic survey.

1. Gender:
   - o Male
   - o Female
2. Age: _____
3. Ethnicity:
   - o White/Caucasian
   - o Hispanic/Latino
   - o Black/African American
   - o Native American
   - o Asian/Pacific Islander
   - o Other
4. What is your cumulative undergraduate GPA? _____ out of 4.0
5. How often do you play video games?
   a. Not at all
   b. About once a month
   c. A few times a month
   d. A few times a week
   e. Every day, but for less than 1 hour
   f. Every day, for 1–3 hours
   g. Every day, for more than 3 hours
6. Do you prefer playing games to other activities (e.g., going out with friends, watching TV)?
   a. Never
   b. Seldom
   c. Sometimes
   d. Frequently
   e. Often
7. Do you consider yourself:
   a. A non-video game player
   b. A novice video game player
   c. An occasional video game player
   d. A frequent video game player
   e. An expert video game player
8. Have you ever played any of the following games for more than a total of <u>two hours</u>?

   Yes No

| Crayon Physics Deluxe, Touch Physics, Apparatus, Magic Pen |
|---|

**Appendix B. Game Play Questionnaire in Physics Playground.**

Direction: On a scale from 1 to 5, how much do you agree with the following statements describing your experience with playing Newton's Playground?

- 1 = Strongly disagree
- 2 = Disagree
- 3 = Neutral
- 4 = Agree
- 5 = Strongly agree

| Question | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Enjoyment<br>1. I enjoyed playing Newton's Playground very much.<br>2. Playing Newton's Playground was fun.<br>3. I thought Newton's Playground was boring (R).<br>4. I would play Newton's Playground in my spare time.<br>5. I would play Newton's Playground longer if I could.<br>6. I would recommend Newton's Playground to my friends. | | | | | |
| Autonomy<br>7. The game provides me with interesting options and choices.<br>8. The game lets me do interesting things.<br>9. I experienced a lot of freedom in the game. | | | | | |
| Competence<br>10. I feel competent at the game.<br>11. I feel very capable and effective when playing.<br>12. My ability to play the game is well matched with the game's challenges. | | | | | |
| Intuitive controls<br>13. Learning the game controls was easy.<br>14. The game controls were intuitive.<br>15. When I wanted to do something in the game, it was easy to remember the game controls. | | | | | |

*Note*: (R) indicates reverse-coding.

## References


Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. New York, NY: Oxford University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). Focus article: how task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research and Perspectives, 12*, 1–33.

Baker, E. L., & Delacruz, G. C. (2008). A framework for the assessment of learning games. In H. F. O'neil, & R. S. Perez (Eds.), *Computer games and team and individual learning* (pp. 21–37). Oxford, UK: Elsevier.

Boyle, E., & Connolly, T. (2008). Games for learning: does gender make a difference. In *Proceedings of the 2nd European conference on games based learning* (pp. 69–75). Reading, UK: Academic Publishing Limited.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258.

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM, 50*(4), 31–34.

Cheung, M. W. L. (2009). Constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(2), 267–294. http://dx.doi.org/10.1080/10705510902751291.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37.

Green, C. S., & Bavelier, D. (2012). Learning, attentional control, and action video games. *Current Biology, 22*(6), 197–206. http://dx.doi.org/10.1016/j.cub.2012.02.012.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westpost, CT: ACE/Praeger.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 169–203). Amsterdam, The Netherlands: Elsevier.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 1–22.

Klimmt, C., Blake, C., Hefner, D., Vorderer, P., & Roth, C. (2009). Player performance, satisfaction, and video game enjoyment. In S. Natkin, & J. Dupire (Eds.), *Proceedings of the 8th international conference on entertainment computing (ICEC 2009) Lecture notes in computer science 5709* (pp. 1–12). Berlin, Germany: Springer.

Klimmt, C., Hartmann, T., & Frey, A. (2007). Effectance and control as determinants of video game enjoyment. *CyberPsychology & Behavior, 10*(6), 845–848. http://dx.doi.org/10.1089/cpb.2007.9942.

Lucas, K., & Sherry, J. L. (2004). Sex differences in video game play: a communication-based explanation. *Communication Research, 31*(5), 499–523. http://dx.doi.org/10.1177/0093650204267930.

Magerko, B., Heeter, C., & Medler, B. (2010). Different strokes for different folks: tapping into the hidden potential of serious games. *Gaming and Cognition: Theories and Practice from the Learning Sciences*, 255–280.

Masson, M. E. J., Bub, D. N., & Lalonde, C. E. (2011). Video-game training and naïve reasoning about object motion. *Applied Cognitive Psychology, 25*(1), 166–173. http://dx.doi.org/10.1002/acp.1658.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. http://dx.doi.org/10.1037/0003-066x.50.9.741.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 59–81). New York, NY: Springer.

Mislevy, R. J., Oranje, A., Bauer, M., von Davier, A. A., Hao, J., Corrigan, S., et al. (2014). *Psychometric considerations in game-based assessment*. New York, NY: Institute of Play.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment. National research council's committee on the foundations of assessment*. Washington, DC: National Academy Press.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2010). *Generation M2: Media in the lives of 8-to 18-year-olds*. Menlo Park, CA: The Henry J. Kaiser Family Foundation. Retrieved from http://kff.org/other/poll-finding/report-generation-m2-media-in-the-lives/.

Riener, C., Proffitt, D., & Salthouse, T. (2005). A psychometric approach to intuitive physics. *Psychonomic Bulletin & Review, 12*(4), 740–745. http://dx.doi.org/10.3758/bf03196766.

Rouse, R., III (2005). *Game design: Theory and practice*. Plano, TX: Wordware Publishing Inc.

Ryan, R., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: a self-determination theory approach. *Motivation and Emotion, 30*(4), 344–360. http://dx.doi.org/10.1007/s11031-006-9051-8.

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, NY: Springer-Verlag.

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. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: Theories and applications* (pp. 295–321). Philadelphia, PA: Routledge/LEA.

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.

Vorderer, P., Hartmann, T., & Klimmt, C. (2003). *Explaining the enjoyment of playing video games: The role of competition* (Paper presented at the Second International Conference on Entertainment Computing, Pittsburgh, PA).

Warren, L. (2009, September 18). *What do we mean when we say non-linear?* [Web log post]. Retreived from http://digitalkicks.wordpress.com/2009/09/18/what-do-we-mean-when-we-say-non-linear/.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: a reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377–392. http://dx.doi.org/10.1177/0734282911406668.

Zapata-Rivera, D., & Bauer, M. (2012). Exploring the role of games in educational assessment. In M. Mayrath, J. Clarke-Midura, D. Robinson, & G. Shraw (Eds.), *Technology-based assessments for twenty-first-century skills: Theoretical and practical implications from modern research* (pp. 147–169). Charlotte, NC: Information Age Publishing.