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

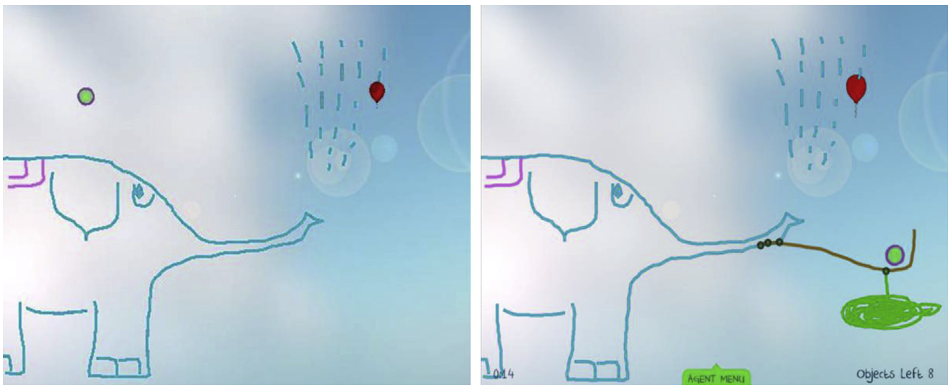

Fig. 1. Trunk Slide (on the left) and a springboard solution for the level (on the right).