Computers in Human Behavior 116 (2021) 106647

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Stealth assessment of creativity in a physics video game

Valerie J. Shute [*], Seyedahmad Rahimi

Florida State University, USA

ARTICLE INFO

Keywords:
Creativity
Stealth assessment
Video games
21st-century skills
Physics playground

ABSTRACT

Creativity has been of research interest to psychologists dating back many decades, and is currently recognized as one of the essential skills needed to succeed in our complex, interconnected world. One medium that has affordances to assess and support creativity in young people is video games. In this paper, we briefly discuss the literature on video games and creativity and provide an example of current work being done relative to measuring creativity in the context of a game called *Physics Playground* using stealth assessment. To validate the stealth assessment of creativity, we conducted a one-group pretest-posttest study with 167 8th and 9th graders from a K-12 school in Florida. Results suggest that our stealth assessment of creativity is valid (i.e., our stealth assessment estimate significantly correlated with our external performance-based measures of creativity). Additional analyses revealed that creativity (i.e., estimated using our stealth assessment of creativity) significantly predicts in-game performance (e.g., number of levels solved), game enjoyment, and learning of physics content. We conclude with a discussion of future directions in this line of creativity research.

1. Introduction

Most of us know creativity when we see it. For instance, consider a group of people living in a remote village in Africa with no electricity or the financial means to buy washing machines. A product design student, Richard Hewitt, came up with a creative idea to couple a bicycle and a large container. Then, with just a bit of re-engineering, he had the *SpinCycle Washing Machine* (Hewitt, 2012; see Fig. 1). This idea came to Richard's mind during a visit to Burundi in Central Africa, after washing about 30 loads of clothes by hand.

Going beyond the final product (like a piece of art, musical score, or clever invention), we want to focus on the *processes* that provide evidence for creativity. In line with this aim, we define creativity as the abilities needed to produce ideas or solutions that are novel yet appropriate for the problem at hand. Moreover, we argue that well-designed games provide an excellent vehicle for capturing and analyzing these processes that can evolve into creative solutions.

Creativity has been of research interest to psychologists dating back many decades, and is currently recognized as one of the essential skills needed to succeed in our complex, interconnected world (e.g., the Partnership for 21st Century Learning, 2019). That is, we are living in a creative society where one's success is based on the ability to think and act creatively. However, despite the recognized importance of creativity, current school systems do not adequately prepare younger people to become creative thinkers (Sawyer, 2011).

As mentioned, one medium with affordances to assess and support creativity in young people is video games. Playing video games is one of the most popular activities for people of all ages. According to the Global Games Market Report (Wijman, 2019) there are more than 2.5 billion gamers across the world. Another recent report analyzed responses to gameplay-related questions from 4500 gamers (ages 18 and older) residing in nine countries (i.e., France, Germany, India, Italy, Japan, Singapore, South Korea, the United Kingdom, and the United States). On average, gamers spend 7 h and 7 min playing video games per week (The State of Online Gaming, 2019). A study on media usage in the U.S. reported that 67% of youth (ages 8 to 18) spend an average of 73 min daily playing video games, compared with only 38 min daily reading print materials (Rideout, Foehr, & Roberts, 2010). Another indicator of the popularity of gaming is that annually, Americans spend $43.4 billion on video/game-related purchases (Entertainment Software Association, 2019); and globally, people spend about $150 billion on games (Wijman, 2019).

So, how can video games cultivate creativity? Will Wright (2006), a renowned game designer, argues that video games are "dream machines" that have the ability to unleash human imagination. He explains that a game is a "possibility space" in which video games start at a

[*] Corresponding author. 3205G Stone Building, Department of Educational Psychology and Learning Systems, College of Education, Florida State University, 1114 W. Call Street, Tallahassee, FL, 32306, USA.
E-mail address: vshute@fsu.edu (V.J. Shute).

https://doi.org/10.1016/j.chb.2020.106647
Received 30 December 2019; Received in revised form 18 October 2020; Accepted 2 December 2020
Available online 9 December 2020

well-defined state and end when a specific state is reached. How players reach a specific goal is open-ended, and each player can navigate this possibility space by making continuous choices and actions.

Gee (2005) similarly describes how a well-designed game incorporates good learning principles that can support players' creativity. First, players are not simply consumers of the game but producers by making their own actions and choices. At a fundamental level, what players do and create in the game to progress through levels is a form of production. For example, in the popular "god" game called *Spore,* players create their own species, and then the species evolve into more intelligent creatures and civilization. Some games, such as *LittleBigPlanet, Portal 2,* and *Physics Playground* have built-in level editor functionality that allows players to modify the games and even create their own levels. Second, good games often encourage players to take risks, explore and try new things, and learn by failing. Failing is not a bad thing in games as it is in traditional education. In fact, failing is a great way to get feedback about progress. Third, video games are "pleasantly frustrating." That is, tasks in a well-designed game are challenging, but reside within a range of difficulty levels. This gives players a great sense of accomplishment upon completing the task.

Despite this inherent link between creativity and video games, there is limited and rather mixed evidence for relationships between playing video games and creativity. For example, Hamlen (2009) investigated the relationship between self-reported time spent playing video games per week and performance on the Torrance Tests of Creative Thinking (TTCT; Torrance, 1972) in 4th and 5th graders. She reported that the number of hours of gameplay does not significantly predict TTCT performance controlling for gender and grade. In contrast, Jackson and colleagues (Jackson et al., 2012) investigated the relationship between gameplay time (i.e., participants' response to *how often do you play videogames?)* and creativity using the TTCT, and they reported that playing video games is significantly associated with creativity.

Although investigating relationships between video gameplay and creativity may be interesting, this line of research does not directly help educators and practitioners to use video games to foster creativity. That is, studies that have investigated the relationship between playing video games and creativity in general (using correlational analyses) are often based on the assumption that creativity is a "general" construct, and do not consider the possible interplay with or dependence on domains. However, Baer and Kaufman (2005), using their Amusement Park Theory (APT) of creativity, suggest that for any creative work to happen, there are some requirements that need to be present. For example, a person working within a particular domain must have at least some basic knowledge about that domain before creative work may emerge. Similarly, Csikszentmihalyi's (1997) theory of creativity includes domain (e.g., math, arts, science, etc.) as one of the components of the creativity model, and creativity occurs at the intersection of person, domain, and field (i.e., experts in the field who can judge the creative work).

Another problem with some existing studies on games and creativity is that they do not clearly state how creativity is defined (see Plucker, Beghetto, & Dow, 2004 for more on this issue). Also, the way in which creativity is assessed in these studies is problematic, where many studies view creativity as unidimensional. Moreover, correlational studies do not systematically examine how specific aspects of creativity manifest in video games. Finally, such general studies (e.g., Hamlen, 2009; Jackson et al., 2012) typically treat all game genres as equal, but that can be misleading as some genres have more potential to enhance creativity than others.

There are, however, some studies reported in the literature (using experimental and quasi-experimental research design) that do use games that are domain relevant and/or have potential for enhancing creativity (e.g., Minecraft, Portal 2). These games allow players to co-create the gaming environment (using the game's create mode or level editor); hence, they permit players to be creative rather than passively receiving game problems to solve. For instance, Fessakis and Lappas (2013) used a physics-related puzzle game called *Crayon Physics Deluxe* (Kloonigames, 2014) to investigate the effects of this game on students' creativity. Similarly, Moffat, Crombie, and Shabalina (2017), and Inchamnan, Wyeth, and Johnson (2013) used another popular game called *Portal 2* (with a high potential to enhance creativity) to investigate the effects of playing this game on participants' creativity, compared to playing two other puzzle games that were had low potential for enhancing creativity (i.e., *I-Fluid*, and *Braid*). Blanco-Herrera, Gentile, and Rokkum (2019) used *Minecraft* (another game with high potential to enhance creativity) to compare its effects on creativity compared to a racing game called *NASCAR*, and watching a TV show. The findings of all of the aforementioned studies showed positive results indicating that certain games (e.g., *Crayon Physics Deluxe*, *Portal 2*, *Minecraft*) that engage players in solving interesting problem, creating virtual environments in the games, and designing new game levels, can enhance people's creativity compared to other video games (i.e., racing or shooting games)—see (in press) Rahimi & Shute, for a full review on the effects of videos games on creativity.

To support creativity using video games in the broader education community, we need to understand the affordances of video games in relation to the multidimensional aspects of creativity. That is, the first question we should ask is: What are some of the cognitive and noncognitive dimensions of creativity that are part of playing video games? In addition, attention needs to be paid to assessment methods that use creative behaviors and products that players create in and outside of video games (Plucker & Makel, 2010). Such behaviors and products are believed to be more valid indicators of creativity than commonly used self-report measures of creativity (McClelland, 1973; Shute, Ventura, & Kim, 2013).

The purpose of this paper is to threefold: (1) review the current literature of creativity and link the literature with the mechanics of popular games that foster creative endeavors; (2) describe a

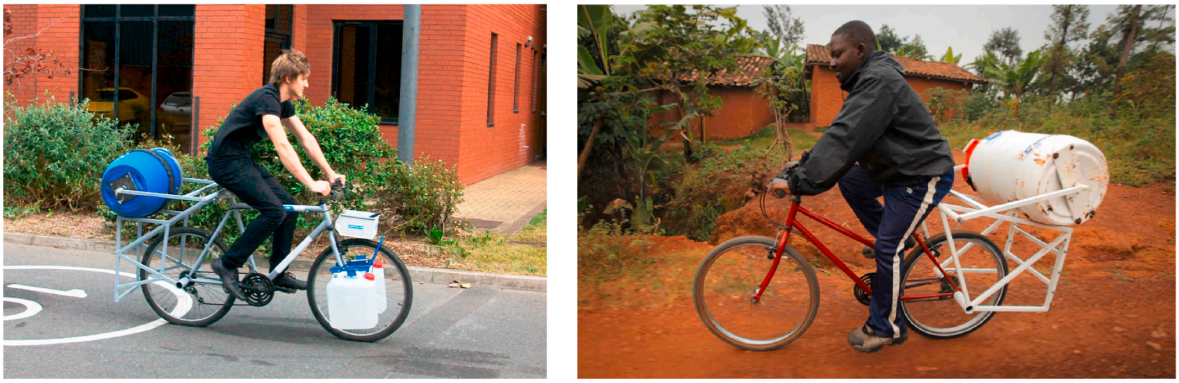

Fig. 1. *SpinCycle Washing Machine* invented by Richard Hewitt (on the left).

methodology called stealth assessment as a way to assess creativity in the context of a learning video game called *Physics Playground* (Shute & Ventura, 2013); and (3) provide empirical support for the construct and criterion validity of our stealth assessment measure of creativity.

2. Review of creativity and the video games literature

2.1. Multiple Dimensions of Creativity

There have been countless arguments over the accepted definition of creativity among psychologists across the decades. Despite this lack of agreement, there are some common notions of creativity that run through the literature. First, creativity is generally defined as the ability to produce solutions, ideas, or products that are both novel and effective (Lubart, 1994). Kaufman and Sternberg (2007) similarly have noted that most definitions of creativity consist of three components: novelty, quality, and relevance.

Second, most research on creativity (e.g., confluence approaches) suggests that there are multiple variables that need to converge for creativity to manifest (e.g., Amabile, 1983; Amabile & Pratt, 2016; Csikszentmihalyi, 1997; Sternberg & Lubart, 1996). For instance, Amabile (1983) emphasized the importance of social and environmental influences on creativity. She noted that creativity is best conceptualized not as a personality trait or a general ability, but instead as a *behavior* resulting from particular collections of personal characteristics, cognitive abilities, and social environments. Similarly, Sternberg and Lubart (1992) explained that the different approaches to creativity could be viewed as a continuum between less contextualized approaches that focus on personal characteristics, and more contextualized approaches that include social-cultural variables that influence individuals' creativity. McCrae (1987) stressed that the ability to think creatively in conjunction with an inclination to do so (i.e., disposition) leads to creative productions.

Another popular theory of creativity was offered by Guilford in his theory of creativity (1956). Although his theory views creativity solely in terms of cognitive abilities, it is multidimensional, operationalized as divergent thinking with four facets: *flexibility* (e.g., the number of categories or themes used when solving a problem or the ability to come up with relevant ideas from different categories or themes); *fluency* (the ability to produce a large number of relevant ideas); *originality* (the ability to produce ideas that are statistically rare); and *elaboration* (the ability to implement and expand on an idea in detail and high quality). For decades, this operationalization has helped researchers design creativity assessments in various environments, with items targeting each particular facet of creativity.

Building on Guilford's work, Torrance suggested that creativity is an everyday phenomenon rather than an unreachable state that only geniuses can achieve (Torrance, 1993). He developed the Torrance Tests of Creative Thinking (TTCT; Torrance, 1972), still used today by many creativity researchers. The TTCT includes both figural and verbal items that assess participants' divergent thinking skills, and many researchers (e.g., Runco & Acar, 2012) view divergent thinking tests as indicators of creative potential (i.e., divergent thinking as proxy for creative thinking). Similar to Torrance, Richards (1990) defines *everyday creativity* as something that anyone can use when dealing with ordinary, day-to-day problems. One area in which everyday creativity frequently happens is when people play certain video games.

In our current study, we focus on three aspects of Guilford's theory of creativity for creating our in-game assessment of creativity—i.e., flexibility, fluency, and originality[1]—in the context of a physics game. There is some research that has investigated creativity specifically in the context of physics education (e.g., Barojas & Pérez, 2001; Cheng, 2004; Diakidoy & Constantinou, 2001). For example, Diakidoy and Constantinou (2001) investigated the relationship of creativity in the domain of physics to the fluency of one's responses to questions, and the type of task administered (i.e., provide an explanation, make a prediction, and apply a concept). An example explanation type of question included: "When I think of iron rusting, wood rotting, and rubber disintegrating, then I am led to believe that 'any material that is taken from nature, with time strives to return to its natural form and environment.' Why might this be happening?" (p. 404). This type of creativity is heavily reliant on formal physics knowledge, which is not the focus of our study. In our study, we focus on conceptual physics understanding and everyday creativity rather than formal physics knowledge and creativity.

In addition to the more cognitive aspects of creativity, there are also noncognitive aspects. For instance, openness to experience, one of the dimensions of the Big-Five factors, refers to a dispositional attribute that is characterized by an awareness of personal feelings and beliefs, receptivity to novel ideas, liberal values, intellectual curiosity, and fantasy (Berzonsky & Sullivan, 1992). Therefore, individuals with higher degrees of openness are described as imaginative, sensitive to aesthetics, curious, independent thinkers, and amenable to new ideas, experiences, and unconventional views (Costa & McCrae, 1992). A long line of research has supported the association between openness to experience and creativity or some aspects of creativity (Costa & McCrae, 1992; Feist, 1999; McCrae, 1987, 1996). For example, McCrae (1987) reported a significant association ($r = 0.40$) between divergent thinking and openness to experience.

Another noncognitive aspect of creativity is the willingness to take risks (i.e., risk propensity), defined as the extent to which an individual takes action knowing there is uncertainty related to the potential pay-off of the action (Dewett, 2007). Risk-taking is associated with openness to change and new ideas (Madjar, Greenberg, & Chen, 2011). According to the literature, willingness to take risks (and knowing the possibility of failing) has been recognized as an essential characteristic of eminent scientists and artists throughout history (Csikszentmihalyi, 1997; Sternberg & Lubart, 1996). For example, Sternberg and Lubart (1992) describe creative individuals as those who "buy low and sell high." They further argue that willingness to take risks is a prerequisite for growth and creativity because one needs to go beyond what is commonly accepted, and learn from various failings. Several studies have reported a positive association between willingness to take risks and creativity (Glover, 1977; Glover & Sautter, 1977). For example, Glover and Sautter (1977) reported that willingness to take risks was significantly correlated with flexibility and originality. Willingness to take risks has also been studied in the context of organizational innovation for many years (e.g., Dewett, 2007; Kogan & Wallach, 1964; MacCrimmon & Wehrung, 1990). For instance, Madjar et al. (2011) found that willingness to take risks is a significant contributor to individuals' creativity and innovation within 12 advertising organizations in Bulgaria.

2.2. Sources of evidence for creativity in video games

The current literature on creativity generally suggests that creativity may be judged by the *output* of creative processes, characterized by both novelty and relevance. Moreover, the creative *process* represents a confluence of factors including personality traits, attitudes, cognitive abilities, knowledge, and the environment. Finally, creativity can be assessed at multiple levels—e.g., the Four-C model of creativity by Kaufman and Beghetto, (2013). To assess people's creativity development in video games, therefore, one needs to consider those aspects of creativity in relation to different sources of evidence that video games can afford. We now provide an example of current work being done relative to measuring creativity in the context of a game called *Physics Playground* (Shute, Almond, & Rahimi, 2019), accomplished by some technology called stealth assessment.

[1] Because the facet of elaboration generally overlaps with the other facets (and we could not ascertain unique indicators for it), we excluded it from our model.

Stealth assessment involves the design, development, and weaving of assessments directly and invisibly into the fabric of any complex learning environment, particularly video games (Shute, 2011; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). During gameplay, players produce rich sequences of actions as the products of continuous interactions with complex tasks. In stealth assessment, the evidence needed to assess targeted skills is provided by the players' interactions with the game itself (i.e., the processes of play).

Inferences on competency states are stored in a dynamic model of the learner (at various grain sizes and in real time). This contrasts with a typically singular outcome of an activity—the norm in educational environments. Stealth assessment may be used to support learning and maintain flow, defined as a state of optimal experience, where a person is so engaged in the activity at hand that self-consciousness disappears, sense of time is lost, and the person engages in complex, goal-directed activity not for external rewards, but simply for the exhilaration of doing (Csikszentmihalyi, 1997). Real-time diagnostic assessment of knowledge, skills, and other attributes of learners can be used to adapt the digital learning environment at hand (e.g., an educational game) to players' needs, such as their current competency level. For example, based on learners' current competency estimates from their in-game performances, the game can adjust task difficulty to levels appropriate to the learners (Kanar & Bell, 2013; Sampayo-Vargas, Cope, He, & Byrne, 2013). Moreover, based on valid inferences, timely and individualized feedback can be presented to enhance learning (Cheng, Lin, & She, 2015; Gobert, Sao Pedro, Raziuddin, & Baker, 2013), especially to support struggling learners (Baker, Clarke-Midura, & Ocumpaugh, 2016). When using such adaptivity in games, flow can be maintained, and learning can occur because tasks are neither too hard/frustrating nor too easy/boring.

Over the past decade, several researchers have created and validated stealth assessments in different games to measure various competencies. Some examples include the following: *problem solving skills* (Akram et al., 2018; Shute, Wang, Greiff, Zhao, & Moore, 2016), *mathematics* (Ke, Parajuli, & Smith, 2019; Ke & Shute, 2015), *computational thinking* (Min et al., 2015), *physics* (in pressShute et al.), *social skills* (DeRosier, Craig, & Sanchez, 2012), and *conscientiousness* (Shute & Ventura, 2013). These efforts have shown that stealth assessment can be used as a reliable and valid assessment method for accurately assessing a range of competencies and attributes.

New developments in psychometric techniques and cognitive theories have enabled the development of stealth assessment—emphasizing the nature of educational assessment as an evidentiary argument. A core element of stealth assessment is the evidence-centered design framework (ECD; Mislevy, Steinberg, & Almond, 2003) that formalizes assessment arguments relative to claims about the learner and the evidence that supports those claims. ECD is flexible enough to reduce constraints of conventional assessment, and it allows for continuous performance data from interactions with complex and interactive environments. An overview of the ECD approach is described in the Method section.

2.3. Stealth assessment of creativity in Physics Playground

Physics Playground (*PP*) is a computer-based game designed to assess and support students' conceptual understanding of physics principles. In *PP*, players draw various objects on the screen using a mouse or stylus, and once drawn, these objects become "alive" and interact with other objects. By playing *PP*, students improve their conceptual understanding of how the physical world operates and how physical objects interact—under the laws of physics.

The game is characterized by an implicit representation of Newton's three laws of force and motion including concepts such as balance, mass, gravity, and conservation of energy and momentum (Shute et al., 2013). These physics principles are operationalized by the use of simple machine-like devices called *agents of force and motion* including ramps, levers, pendulums, and springboards to lead a green ball to the red balloon on the screen. Most of the 100s of levels in *PP* can be solved by various solutions, using more than one agent. Thus, *PP* allows players to be creative and produce interesting mechanical devices to solve levels, some of which even the designers of the game did not expect. Furthermore, players often attempt multiple times to achieve the "most awesome" solution.

To assess these creative behaviors in the game, we identified three creativity competency model variables—fluency, flexibility, and originality, and identified in-game observables that provide evidence for those variables (i.e., evidence model variables). Table 1 summarizes the creativity competency and evidence model variables (i.e., indicators) in *PP*.

Here is an illustration of how these variables work to assess players' creativity in the context of *PP*. Fig. 2a shows how a level called *Big Watermill* looks like when the level starts. The ball falls down due to gravity, goes out of the screen, then reappears in its initial position. This loop continues until the player draws something to redirect the ball to the balloon. The most common solution among *PP* players (and expected by the designers) involves drawing a ramp to direct and propel the ball to the balloon from underneath the counter-clockwise rotating watermill (shown in Fig. 2b). Any trajectory of the ball—leading to solving the level—that deviates from the trajectory shown in Fig. 2b can provide evidence for originality as it is likely to be a rare (thus novel) solution. Again, there is only a brief amount of time for the player to catch the ball from its original position and direct it to the balloon any other way other than from under the watermill. Only a few players (out of 100s) directed the ball to move *above* the watermill—a solution similar to Fig. 2c. In this case, the player created a lever, trapped the ball inside a circular area on the left side of the lever, created a weight on the right side of the lever, and used the watermill itself as a fulcrum. The weighted lever served to pull down and direct the ball to the balloon. Such a solution provides positive evidence for originality in *PP*. That is, the trajectory of the ball (shown as dotted line) in this solution (2c) deviates from the common solution, and the unexpected agent employed (lever) provides evidence for *originality*.

Specific indicators are automatically identified and scored during gameplay. For example, in *PP*, the game engine tracks the trajectories of the players' ball in a successful solution (i.e., the set of x, y coordinates), and saves them out as series of vector values in the log file. Those vector values can then be compared to the most common trajectory, and large differences between trajectories are thus evidence for originality.

Establishing these evidence model variables (also known as indicators) and scoring rules to decide *when* those indicators provide evidence for creativity can be tricky depending on the nature of a given level. Furthermore, as "gaming the system" is not always viewed negatively in the gaming context, differentiating creative solutions from solutions that exploit the features of the game is critical (Kücklich, 2004).

Table 1
Competency and evidence model variables for creativity assessment in physics playground.

CM variables	EM variables
Fluency	- Number of all agents drawn per solved level - Number of all agents drawn per unsolved level - Number of objects drawn per solved level - Number of objects drawn per unsolved level
Flexibility	- Number of applicable agents attempted in the level - Standard deviation among agent frequencies of [R] - Consecutive use of incorrect agent [R]
Originality	- Difference between ball trajectory in a solution from expected trajectory - Use of an unexpected agent to solve the level

Note. *R indicates reverse coding (for negative evidence).

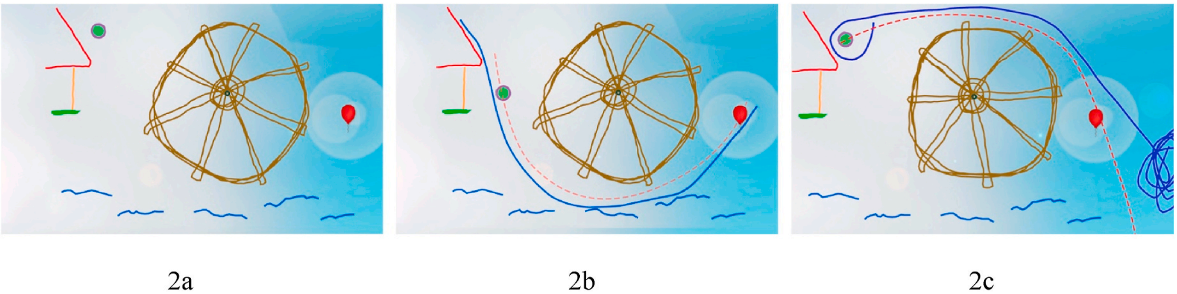

Fig. 2. The *Big Watermill level* (2a) and two possible solutions (2b and 2c).

As the very definition of creativity emphasizes both *novelty* and *relevance*, in-game behaviors that are *not* appropriate in terms of the rules and mechanics of the game should not be considered as evidence for creativity. For instance, Fig. 3a shows a level called *Need Fulcrum.* As the level starts, the ball lands on the horizontal blue line towards the left of the screen. Players are expected to create a lever (as shown in Fig. 3b) and drop a weight on it, to launch the ball toward the balloon. The game level's name, *Need Fulcrum*, provides a hint that drawing a lever is a reasonable solution option. A number of the players, however, figured out that if they quickly drew lines under the ball, they could solve this level without using any agents of force and motion (shown in Fig. 3c). Although the trajectory of the ball shown in Fig. 3c deviates from the one in Fig. 3b, such a solution does not provide evidence for *originality* as this solution violates the rules of the game (i.e., no simple machine was applied to solve the level). Furthermore, the "solution" was not original in that many students actually tried this hack. Stealth assessment identifies in-game behaviors specific to the levels of the game.

To test whether we could actually measure creativity via stealth assessment in *PP*, we conducted a study with the following research questions:

1) Is our stealth assessment estimate of creativity valid (i.e., does it correlate with other external measures of creativity)?
2) Does creativity predict in-game performance (i.e., number of levels solved, number of gold and silver coins earned)?
3) Does creativity predict enjoyment of the game?
4) Does creativity predict physics learning?

Our hypotheses behind these research questions, based on the literature, are as follows. Relating to research question 1, we expected that our stealth assessment estimate of creativity would show evidence of *construct validity* - defined as the degree to which something measures what it purports to measure. Thus, we hypothesized that our creativity measure would significantly correlate with the external measures of creativity as we used both a theory-driven and performance-based method of assessment. In particular, because we focused on divergent thinking skills in our competency model (fluency, flexibility, and originality) and not dispositional variables (like the openness survey measures), we expected higher correlations among the former than the latter.

We also predicted that our stealth assessment estimate of creativity would impact various outcome measures (i.e., game performance, enjoyment, and learning physics), reflected in research questions 2-4. This refers to tests of *criterion validity* - i.e., the degree to which something can predictively or concurrently measure something. We know from the literature that creative players tend to be good at divergent thinking—they can come up with various ideas and solutions while solving different problems (e.g., McCrae, 1987). Therefore, the odds of players getting stuck in the game are lower for more creative than less creative players. Hence, regarding our game-performance outcomes, creative players should be able to solve more levels, receive more coins (gold and silver), and generally make more progress in the game compared to less creative players.

In terms of game enjoyment, research has shown (e.g., Amabile, 1983) that making progress in meaningful tasks is essential to boosting individuals' positive affective states and eventually their creativity. When creative players perform well in the game, they should report a higher level of game enjoyment compared to less creative players (Gee, 2005; Velikovsky, 2014). A higher level of game enjoyment by more creative players can also be viewed via flow theory (Csikszentmihalyi, 1997). Specifically, Velikovsky (2014) asserts that "flow theory in creativity can equate to the 'fun factor' in games" (p. 8).

Finally, our hypothesis about creativity and learning is that because creative players are likely to perform better in the game, they may learn more of the underlying physics content compared to less creative players. We know from prior studies with *Physics Playground* that how a player performs in the game directly effects their degree of learning (e.g., Shute et al., 2015). Therefore, when more creative players perform better in the game, they should end up learning more compared to less creative players.

3. Method

3.1. Participants and research design

The participants of this study consisted of 167 8th and 9th graders (76 male and 91 female; 13-15 years old) from a K-12 school in Florida. Upon the completion of the study, each student received a $25 gift card.

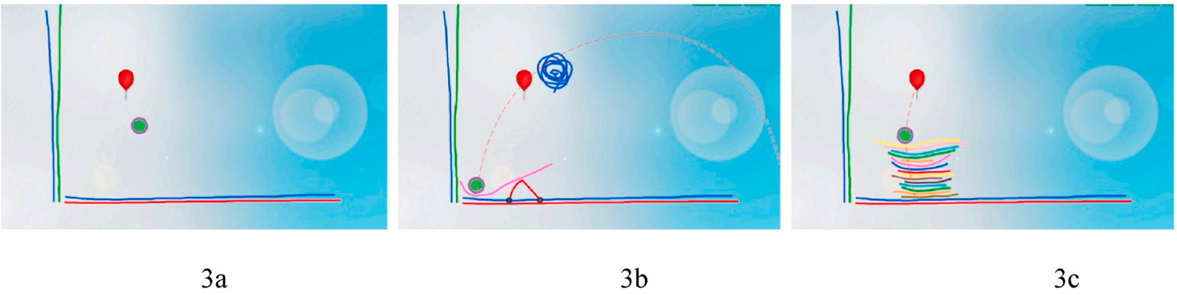

Fig. 3. The *Need Fulcrum* level (3a) and two possible solutions (3b and 3c).

We used a one-group, pretest-posttest research design.

3.2. Procedure

The total gameplay time was about 4 h (across six 45-min sessions in a week). Thirty computers, in one of the school's two computer labs, were used for this study. Separators were used between the computers to make sure that students did not talk to each other during gameplay. All students played the same version of *PP* with 74 levels within seven playgrounds.

After introducing the study, researchers administered a brief online demographic questionnaire about students' age, gender, and grade. Afterwards, an online physics pretest, followed by some measures of creativity (i.e., alternative uses test and openness survey) were administered. Upon completing the pretest battery of measures, the researchers introduced *PP* to the students. To encourage students to pay attention during gameplay, they were told that the student with the most gold coins at the end of the study would receive an extra $25 gift card.

The first session of gameplay started by having students complete the agent-tutorial videos (about 5 min in duration). Students could start playing the game when they finished watching the videos. Researchers instructed students to start by playing levels in Playground 1 (i.e., mostly easy levels), and then they could move on to any level in any playground they wanted (students were informed that the difficulty of levels across the 7 playgrounds progressively increased). Specifically, researchers instructed participants that, "*Your goal is to solve as many of the problems, in as many awesome ways as you can. The tools we taught you will come in handy for many problems. Feel free, however, to solve any problem in whatever way you like.*" These instructions were provided to encourage students to both do well in the game and be creative (i.e., solve levels in as many ways as they could). During gameplay, students were told that they could watch the agent-tutorial videos if they were struggling in a level. Students played the game for 4 out of 6 sessions of the study. Sessions 1, 2, 3, and 5 were gameplay sessions. Sessions 4 and 6 were game design sessions, using the game's level editor (discussed in the next section on Measures). After the last game design session, students completed a posttest of physics.

3.3. Measures

Physics Test. Working with a physics expert, we created 24 multiple-choice items, counterbalanced between two equivalent forms (Form A and Form B) and used for pretest and posttest in the study. Each form included two items for each of our six main physics concepts. The tests measured students' understanding of Newtonian physics (i.e., Newton's 1st and 2nd laws of force and motion, conservation and transfer of momentum, and potential and kinetic energy). The reliability (Cronbach's α) for the physics test Form A was 0.72 and for Form B was 0.73.

Game Enjoyment. After the students completed playing *PP*, we asked them two questions about how they enjoyed playing the game (Cronbach's α = .80) using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The game-enjoyment items included: "I enjoyed playing *Physics Playground*" and "I would play this game in my spare time," We used the average of these two items to compute a single score for game enjoyment.

External Measures of Creativity. To measure students' creativity with an external measure for validation purposes, we used Wallach and Kogan's Creativity Test (Wallach & Kogan, 1965) consisting of three *alternative-uses* test items (e.g., "How many different uses can you list for a rubber band?") with a maximum of eight possible responses. Students had 1 min to compile their lists with as many answers as they could. When time was up, they circled their top two most creative responses, based on the procedure suggested by Silvia (2008a). We were thus able to score responses for fluency (i.e., number of responses per item; Cronbach's $\alpha = 0.90$) and originality (i.e., two most creative responses; Cronbach's $\alpha = 0.80$). The reason for scoring just fluency and originality was because Wallach and Kogan (1965) only scored their test on these two dimensions—i.e., the number and uniqueness of responses. Fluency and flexibility tend to be highly correlated, thus in the literature, researchers have adopted the same scoring standard of focusing on fluency and originality in the Wallach and Kogan test (e.g., Cheung, Lau, Chan, & Wu, 2004; Cropley, 1968; Silvia, 2008b).

Two trained raters independently rated the two most creative responses on originality (using an inverse of frequency) and resolved any disagreements. To determine originality of the circled "most creative" uses, the raters employed a list of common and uncommon responses (e.g., using a rubber band to tie up one's hair or to make a slingshot were common, while using a rubber band to create a gripper for stuck lids was uncommon). In addition, we administered an online *openness to experience* questionnaire (McCrae & Costa, 1985), a 10-item survey with 5-point Likert scale items (e.g., "I enjoy hearing new ideas.") with responses from 1 (strongly disagree) to 5 (strongly agree)—see Appendix A.

Creativity Assessment of Student-Made Levels. During sessions four and six of the study, students used *PP*'s level editor to design their own levels. They were asked to create an awesome level that a good friend of theirs would enjoy playing. Using the scoring rubric shown in Table 2, two trained raters (not the authors) independently scored each level that the students designed in both sessions with interrater reliability of $r = 0.91$. To employ these student-created levels as a creativity measure, we identified features of the levels that represent aspects of creativity. This type of assessment is neither a true external assessment of creativity, nor an in-game assessment of creativity. Instead, it is something in between. However, because it's not part of students' gameplay activities, we are referring to it as an "external" measure. Table 2 describes the creativity dimensions and scoring rules used to make a holistic judgment about the student-made levels.

Note that the creativity dimensions of relevance/appropriateness, originality, and elaboration were discussed earlier in the Multiple Dimensions of Creativity section of this paper. We included two additional creativity dimensions in our rubric - the aesthetics of the creation, and

Table 2
Creativity scoring rubrics for student-made levels in physics playground.

Categories	Scoring rules
Relevance	*Can it be solved?* (This is a screening criterion) o If unsolvable, then don't score other variables = 0 o If solvable = 1
Originality	*Is it original relative to existing levels?* (Possible scores: 0, 1, and 2) o Almost identical to an existing level = 0 o Has some similarities = 1 o Very dissimilar = 2
Aesthetics	*Is it aesthetically pleasing?* (Possible scores: 0, 1, and 2) o Aesthetically unappealing with poor visual elements = 0 o Plain with completed visual elements = 1 o Very pleasant with well-thought-out visual elements = 2
Humor/ Surprise	*Is it humorous or surprising* (i.e., Does it make you smile or did it surprise you)? (Possible scores: 0, 1, and 2) o Not humorous or surprising at all = 0 o Somewhat humorous or surprising = 1 o Very humorous or surprising = 2
Elaboration	*How difficult is it?* (Possible scores: 0, 1, 2, 3, and 4) o If the balloon is located above ball = 1, if not = 0 o If any agent other than ramp is used to solve the level = 1, if not = 0 o If obstacles to remove/avoid are present = 1, if not = 0 o If the ball is falling out of the problem space = 1, if not = 0 *Note*: add the scores for each item to get the elaboration score
Total	Add all the scores to get the creativity score for each student-made level

its humor/surprise. Regarding the aesthetics dimension, Cropley and Cropley (2011) argued that creativity should include aesthetic properties. That is, while beauty or aesthetics is important in creativity, not every beautiful or aesthetically pleasing product is creative. Creativity should exceed ordinary beauty through novelty, unusualness, and appropriateness (Cropley & Cropley, 2011). Runco (2003) further noted that while originality is required for creativity, an original idea or solution might lack an aesthetic appeal that characterizes truly creative ideas. Finally, one of the characteristics that distinguishes creative individuals is aesthetics sensibility (Abdulla & Cramond, 2017; Cropley & Cropley, 2011). Therefore, we included aesthetics as one of the aspects for measuring the creativity of the student-created game levels.

In relation to the humor or surprise aspect of creativity, Jackson and Messick (1965) suggested that there are three criteria for a product to be called creative: unusualness, appropriateness, and transformation. These three criteria should generate three corresponding responses of surprise, satisfaction, and stimulation. The surprise response occurs when a product "catches our eye" and is unusual. The satisfaction reaction happens when the product is appropriate for the context, and satisfaction depends on how well the product meets the expectations for appropriateness. In our rubric, we used relevance as a means to assess the appropriateness of the game levels—any level that can be solved is appropriate. Moreover, we included humor/surprise in our rubric to measure the unusualness of the game level (in terms of being humorous or surprising, thus adding to the level's unusualness).

Based on the scoring rules described in Table 2, the maximum creativity score that a level can receive is 11. Fig. 4 includes some of the student-created levels and associated scores based on the scoring rules. "Derp Invasion" was judged to be a fairly creative level as it is a medium difficulty level (3/4) that is solvable (1/1), very different from the existing levels (2/2), aesthetically pleasing (2/2), and somewhat humorous/surprising (1/2). Although "Hoop City" received the same scores for most categories, it scored lower than "Derp Invasion" as it is an easy problem (i.e., it can be solved by simply drawing a ramp over the basketball). Although "Monkey" could be a fairly creative level, it scored 0 as it is not solvable (the ball is stuck in the left ear of the monkey, and thus is impossible to get it out). "Sunny" similarly was scored low as it received a score of 0 for originality as there is already a level in the game called "Sunny" that looks very similar to the created level.

ECD Models for Our Stealth Assessment of Creativity. The primary purpose of an assessment is to collect information that will enable the assessor to make inferences about students' competency states—what they know, believe, and can do, and to what degree. Accurate inferences of competency states support instructional decisions that can promote learning. ECD defines a framework that consists of three main models that work in concert.

The ECD framework allows/requires an assessor to (a) define the claims to be made about students' competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of tasks that will elicit that evidence. These three actions map directly onto the three main models of ECD shown in Fig. 5.

A good assessment has to elicit behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence in terms that suit the purpose of the assessment. Working out these variables, models, and their interrelationships is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design:

- *What collection of knowledge, skills, and other attributes should be assessed?* (i.e., Competency Model). Variables in the competency model (CM) are usually called "nodes" and describe the set of person variables on which inferences are to be based. The term "student model" is used to denote a student-instantiated version of the CM—like a profile or report card, only at a more refined grain size. Values in the student model express the assessor's current belief about a student's level on each variable within the CM.
- *What behaviors or performances should reveal those constructs?* (i.e., Evidence Model). An evidence model (EM) expresses how the student's interactions with, and responses to a given problem constitute evidence about competency model variables. The EM attempts to

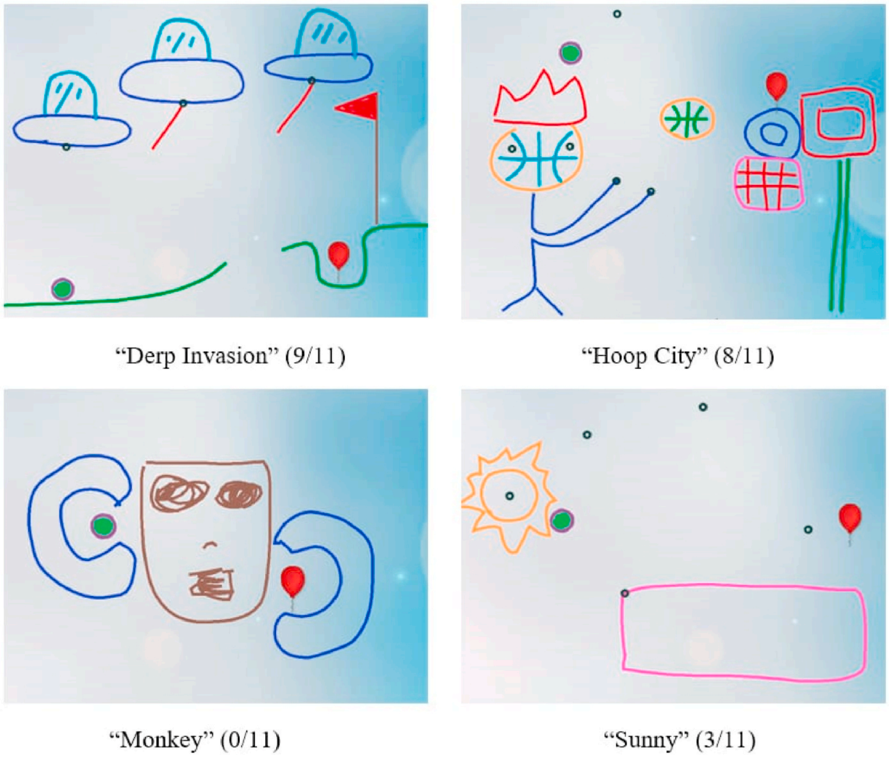
"Derp Invasion" (9/11)
"Hoop City" (8/11)
"Monkey" (0/11)
"Sunny" (3/11)

Fig. 4. Examples of student-made levels and associated scores.

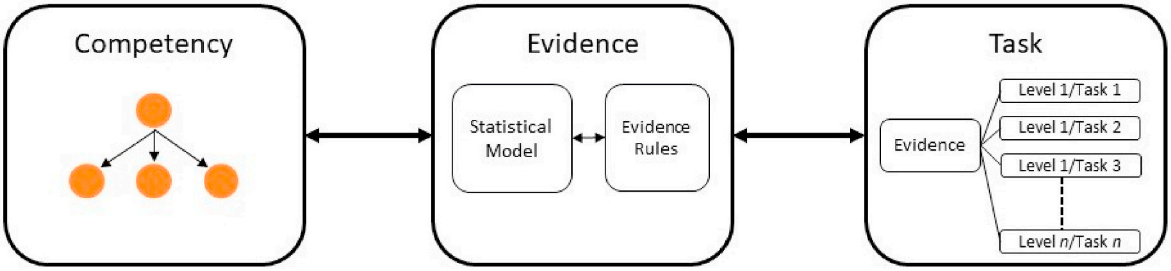

Fig. 5. Three main models of an ECD-based assessment (adapted from Mislevy et al., 2003).

answer two questions: (a) What behaviors or performances reveal targeted competencies, and (b) What's the connection between those behaviors and the CM variable(s)? Basically, an EM lays out the argument about why and how the observations in a given task situation (i.e., student performance data) constitute evidence about the CM variables.

- *What tasks should elicit those behaviors that comprise the evidence?* (i.e., Task Model). A task model (TM) provides a framework for characterizing and constructing situations with which a student will interact to provide evidence about targeted aspects of knowledge or skill related to competencies. These situations are described in terms of (a) the presentation format (e.g., directions, stimuli), (b) the specific work or response products (e.g., answers, work samples), and (c) other variables used to describe key features of tasks (e.g., knowledge type, difficulty level). Thus, task specifications, in an educational context, establish what the student will be asked to do, what kinds of responses are permitted, what types of formats are available, and other considerations, such as whether the student will be timed, allowed to use tools (e.g., calculators, dictionaries), and so forth. Multiple task models can be employed in a given assessment. Tasks are the most obvious part of an assessment, and their main purpose is to elicit evidence (which is observable) about competencies (which are unobservable). In the context of a learning game, game levels are the tasks students need to work on.

In short, the ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about student competencies via an evidentiary chain (e.g., arguments that serve to connect task performance to competency estimates), and are thus valid for their intended purposes.

Stealth Assessment of Creativity. We developed three different stealth assessment measures of creativity—for fluency, flexibility, and originality as shown in Table 1. To estimate *fluency*, we identified different variables such as "Number of drawn objects per solved level." For *flexibility,* we collected data on relevant observables such as the "consecutive use of an incorrect agent" [reverse coded]. To estimate *originality*, in the log files, we captured the solution trajectory in a solved level via the student's x, y coordinates (i.e., the path the ball took from origin to hitting the balloon). We had, for each level, expert (or expected) solutions (with x, y coordinates) thus we could calculate unique student solutions to a level based on the differences between the student and expert solution paths. In this case, large discrepancies between student and expert solutions (relative to the x, y coordinates) were judged as more original (i.e., novel or unexpected) than small discrepancies.

After establishing the relevant observables for each of the three facets of creativity, we created a Bayesian network (BN) for each one of the 74 levels in the game (given that levels differed in terms of difficulty levels and targeted physics content) to estimate students' creativity in *PP,* using Netica (by Norsys Software Corporation). According to a recent review by de Klerk, Veldkamp, and Eggen (2015), BNs are the most frequently-used analytical and data modeling framework to analyze learners' performance data in game-based and simulation-based assessment. Other modeling methods include Confirmatory Factory Analysis, Epistemic Network Analysis, Multidimensional Item Response Theory, Educational Data Mining, and Artificial Neural Networks. There are several advantages to using BNs as a data modeling framework in game-based assessment, like stealth assessment: (1) BNs provide an easy-to-view graphical representation of the competency model (direct and indirect relationships among variables) for clear operationalization; (2) BNs can "learn" from data as they are probability models (thus make probabilistic predictions) thus are improved beyond the original model as more data become available; (3) Updating BNs is immediate (as performance data come from the game environment) compared to other analytical approaches which tend to be post-hoc, so they provide real-time diagnosis—overall and at various grain sizes; and (4) Enhancements to BN software permit large and flexible networks with as many variables as wanted (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). In addition, by using only discrete variables, BNs can be scored very quickly, making them suited for embedded scoring engines.

Our BN system followed the algorithm described in Chapter 13 of Almond et al. (2015). Each level in *PP* corresponds to one problem to be solved (i.e., hitting the balloon using the ball by drawing objects). As shown in Fig. 6, creativity is the parent node of fluency, flexibility, and originality. Each time a student plays a level, he or she generates evidence for the child nodes of creativity. As students play and provide positive or negative evidence for each child node, the parent node (i.e., creativity) is updated using two processes—Evidence Identification (EI) and Evidence Accumulation (EA). That is, at the end of each level (i.e., when a student quits a level or solves it successfully), the log files are automatically parsed, observables (i.e., key features of the player's log data) are identified (i.e., EI process) and scored (using relevant scoring rules, e.g., "if the trajectory of the solution differs by one standard deviation from the expected solution, mark the solution as "rare"), and the scores are absorbed by the relevant level-specific BN for each student (i.e., EA process). The BN is then updated at the end of each level with the current probabilistic estimates of the student's creativity—overall and at the facet level (see Kim, Almond, & Shute, 2016; Shute & Ventura, 2013).

More specifically, the BN automatically calculates the low, medium, and high probabilities for fluency, flexibility, originality, and then for overall creativity per student. These probabilities were calculated using a simple gradient descent algorithm implemented in the CPTTools package (Almond, 2010, 2015). The prior probabilities in the BN shown in Fig. 6 were based on normal probability distributions (i.e., before gameplay data was entered to the BNs) for each indicator node (except for the "deviation from expected trajectory" node) according to our experts' judgments. The reason for having a different prior distribution for the trajectory indicator was because our experts believed that it was more likely to have common than usual and rare trajectories. This distribution is also in alignment with the literature on the distribution for originality (e.g., Cropley, 1972; Dippo, 2013), suggesting that originality usually follows a positively skewed distribution (i.e., more common than rare responses). After each student provides gameplay data per level, these initial probabilities get updated based on the BNs. The next steps of this work could use the current posterior distribution of

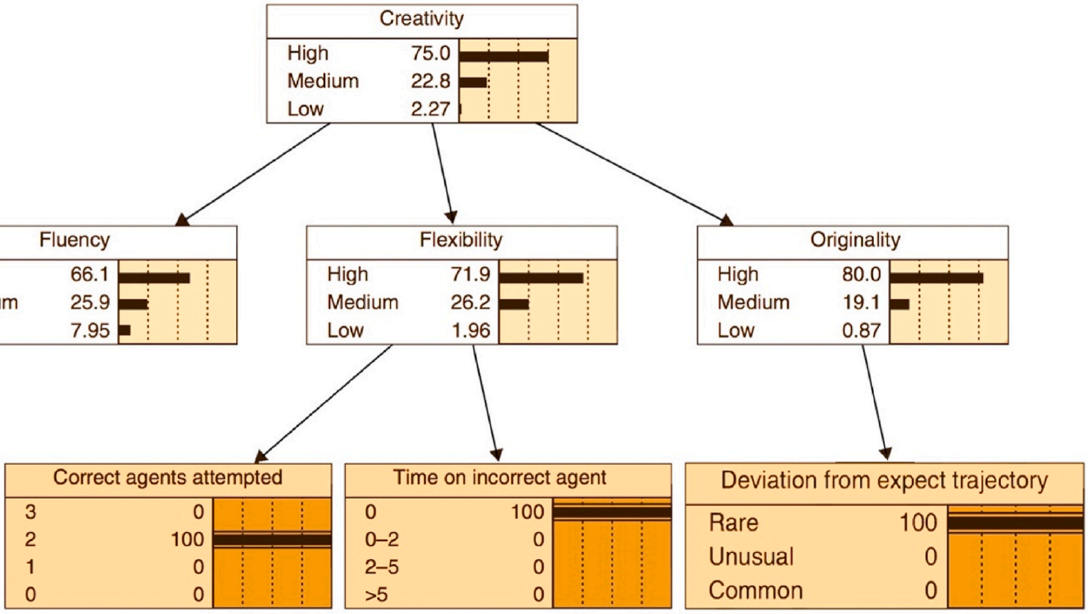

Fig. 6. Creativity BN with its child nodes and example indicators for one level in *PP*.

probabilities as the prior distributions in other studies using Physics Playground.

To understand how the overall creativity scores are estimated, let $X_1, \ldots, X_J$ be the variables in the creativity BN (i.e., fluency, flexibility, and originality). For every variable, X_j, $pa(X_j)$ is the parent of X_j in the graph. Then the joint distribution of the variables is:

$$Pr\left(X_1, \ldots, X_J\right) = \prod_{j=1}^{J} Pr\ \left(X_j|\mathrm{pa}(X_j)\right)$$

In this equation, $\Pr(X_j|pa(X_j))$ is the conditional probability table (CPT) for X_j. Since the creativity BN in this study is monotonic, and its variables are discrete, the conditional probability distribution becomes a conditional probability table. To update the CPTs, two steps are taken: (1) an expected maximization (EM) algorithm is used to find the expected CPTs for all nodes, and (2) a gradient descent algorithm is used to learn the new CPTs for *each* node based on the expected CPTs, and then the new CPTs are inserted into the BN (to read more on this topic see in pressTingir & Almond).

So Fig. 6 shows the trajectory of the solution for a given level that has been scored as "rare." This generated a high probability for originality, and in turn, for overall creativity. Considering other pieces of evidence coming to the BN in this example, the high, medium, and low probabilities are calculated as P(Creativity = high | evidence) = 0.75, P(Creativity = medium | evidence) = 0.23, and P(Creativity = low | evidence) = 0.02.

These estimates become increasingly accurate as more data is absorbed into the BNs at the end of each level. For the purpose of our analyses in the current study, we computed a single value for the parent node (creativity) per student. That is, the stealth assessment estimate consists of three probabilities (i.e., high, medium, and low). We assigned numeric values to the three states and computed the expected value. This Expected A Posteriori (EAP) value can also be expressed as, $P(\theta_{ij} =$ High) - $P(\theta_{ij} =$ Low), where θ_{ij} is the value for Student i on Competency j (e.g., creativity), and [1 × P(High)] + [0 × P(Med)] + [-1 × P(Low)] = P(High) - P(Low). This results in a single value from −1 to 1. For example, the EAP of creativity in Fig. 6 is computed as (1 × 0.75) - (1 × 0.02) = 0.73. We used this value at the end of all gameplay for our analyses (i.e., correlation with external measures and conducting regressions).

In-game Measures of Game Performance. Students' interactions with the game were recorded in the log files. Parsing the log files, we could create many different variables indicating students' performance in the game. Some of these performance measures included the following: (1) *total number of gold coins*: depending on the levels' difficulty, we assigned the minimum number of objects needed to solve the level (i.e., par value, like in golf). When a student solved a level under par, she or he would receive a gold coin—a more elegant or efficient solution; (2) *total number of silver coins*: when a solution used more objects than established by the par to solve the level, the student received a silver coin; and (3) *total number of levels solved*: this represented the total number of levels a student solved across all gameplay sessions.

4. Results

To address research question 1 regarding the criterion-related validity of our stealth assessment estimate of creativity (computed using BNs), we conducted several correlational analyses. As expected, our creativity estimate correlated with our external, performance-based measures of creativity as follows: (a) *Alternative Uses test—fluency* (r = 0.18, p = .02); (b) *Alternative Uses test—originality* (r = 0.18, p = .02); and (c) *Student-made levels* (r = 0.23, p = .01). There was no significant correlation between our creativity estimate and the *Openness survey* (r = 0.02; p = .77). For our external performance-based measures of creativity—Alternate Uses (both fluency and originality), and the Student-made levels—the significant correlations suggest that our stealth assessment measure of creativity is valid.

To answer research question 2, we conducted three separate multiple regression analyses testing whether or not creativity can predict game performance outcomes controlling for the pretest. We set gold coins, silver coins, and levels completed as the dependent variables, and pretest and our creativity estimate as the independent variables in each regression. Results showed that, controlling for incoming knowledge, our creativity estimate was not a significant predictor of gold coins earned ($\beta_{creativity}$ = 0.11, t = 1.38, p = .17; $\beta_{pretest}$ = .31, t = 3.92, p = .17; F(2, 163) = 12.57, p < .001, R^2 = 0.12). However, our creativity estimate significantly predicted silver coins earned ($\beta_{creativity}$ = 0.27, t = 3.29, p = .001; $\beta_{pretest}$ = -.09, t = −1.04, p = .30; F(2, 163) = 5.44, p = .005, R^2 = 0.05), and total number of levels solved ($\beta_{creativity}$ = 0.39, t = 5.20, p < .001; $\beta_{pretest}$ = .16, t = 2.11, p = .04; F(2, 163) = 23.38, p < .001, R^2 = 0.21). In other words, for each one standard deviation (SD)

increase in the creativity estimate, the number of (a) silver coins increases by 0.27 *SD*, and (b) levels solved increases by 0.39 *SD* controlling for the pretest. Thus, the more creative students earned more silver coins, and completed more game levels than less creative students.

For research question 3 regarding the relationship between creativity and enjoyment of the game, we tested the relationship of the game enjoyment score (i.e., the average of the two items related to game enjoyment) and the stealth assessment estimate of creativity using a simple regression. Results showed that our creativity estimate significantly predicted students' game enjoyment ($\beta = .21$, $F(1, 152) = 6.73$, $p = .01$, $R^2 = 0.04$). That is, with one *SD* change in the creativity estimate, students' enjoyment changes by 0.21 *SD*.

Finally, to address research question 4, testing whether our stealth assessment estimate of creativity predicts learning physics from the game, we conducted another simple regression analysis with posttest score as the dependent variable and our in-game creativity estimate as the independent variable. Results showed that our creativity estimate significantly predicted students' posttest scores ($\beta = 0.19$, $F(1, 152) = 5.64$, $p = .02$, $R^2 = 0.04$). That is, with one *SD* change in the creativity estimate, students' posttest scores change by 0.19 *SD*. However, when the physics pretest score was included in the equation, our creativity estimate was no longer a significant predictor of the posttest scores ($\beta = -.06$, $t = -.81$, $p = .42$) with a model R^2 of 0.35.

5. Discussion and future research

As playing video games has become a key part of everyday life for today's youth, the broader education community has been exploring affordances of video games to measure and support competencies that are valuable to success in the 21st Century. In this paper, we discussed how one such game—*Physics Playground*—can be used as a vehicle to measure creativity. We examined research questions related to the validity of our stealth assessment estimate of creativity, as well as the effects of creativity on in-game performance, enjoyment, and learning in *Physics Playground*.

The results of our study showed that the stealth assessment estimate of creativity appears to be valid as it significantly correlated with the performance-based measures of creativity—i.e., the Alternative Uses test (fluency and originality scores) and the student-made levels (overall score of creativity based on scoring rubrics). Therefore, our hypothesis about the validity of our stealth assessment measure was mostly supported. We did not find a significant correlation between our creativity estimate and the openness survey, most likely because people often misrepresent themselves on these types of surveys. That is, there is a tendency for people to answer in line with what society or the researchers view as favorable rather than their actual beliefs. This effect can lead to the inflation of scores related to good behaviors and/or the reduction of reported bad behaviors in the self-report. Another issue with self-report is that people sometimes have different conceptual understanding of the questions. These weaknesses may undermine the reliability and validity of self-report measures as an ideal external assessment.

In general, our preliminary criterion-validity finding is promising as automated, real-time assessment of complex constructs like creativity is difficult to accomplish (Shute & Wang, 2016). This validation was made possible by using stealth assessment, powered by the ECD models, to collect data (indicators) that have been theoretically linked to the construct in question. Analysis of student performance can then be computed in real-time and at various grain sizes (e.g., overall creativity, or at the facet level for more diagnostic information). In contrast, data-driven approaches have been used in game-based assessment, such as data mining and machine-learning techniques, but these methods (e.g., clustering, classification, prediction, and patterns tracking) tend to be bottom up and exploratory, thus missing the theoretical foundation of the construct we wanted to assess. These two methods for understanding large amounts of student performance data are both valuable, but applicable in different situations and for different purposes.

There are other similar measurement techniques that can be used to assess various competencies. For example, choice-based assessment, introduced by Schwartz and Arena (2013), employ short and engaging games (i.e., *choicelets*) that comprise the environment for assessing students' targeted knowledge and skills. Similar to stealth assessment, the choices that a student makes (related to the concept that they are supposed to learn) get logged in the log files and then analyzed. Schwartz and Arena (2013) present an example about assessing critical thinking using choice-based assessment in the context of a game about color mixing. A full comparison of this assessment method and stealth assessment is outside the scope of this paper, but the main difference is that in stealth assessment, three core ECD-based models are used as a framework but in choice-based assessment, no clear assessment framework is mentioned. Another difference is that stealth assessment has been used for complex games rather than short, choice-based games. Again, using an assessment method should be based on the purposes of the assessment, and one assessment method is not necessarily superior to another.

We also found that creativity is important to productive in-game performance, such as attaining silver coins for solutions, as well as solving game levels. This finding was expected, and it makes sense given that the game requires players to physically create solutions—i.e., using a mouse or stylus to draw objects that come alive on the screen when drawn. However, our creativity estimate was not a significant predictor of the gold coins earned. This makes sense, because a gold coin was awarded for a very efficient solution to a level (i.e., using a minimum number of objects drawn). Therefore, students who were thinking about solving levels efficiently received more golds than students who were trying to be creative (have multiple agents and objects in their solutions).

Creativity also predicts enjoyment during gameplay. This finding similarly was foreseen, and aligns with related literature on creativity (e.g., Amabile, 1983), video games (Gee, 2005), and flow (Csikszentmihalyi, 1997; Velikovsky, 2014). That is, because players need to create objects on the screen to solve game levels, more creative players do better in the game than less creative players. Divergent thinking allows them to try various solutions for a given problem (McCrae, 1987), so when they get stuck, they can just try another way to solve the level. As Amabile (1983) noted, making progress in any endeavor can positively boost students' affective state and enhance their creativity. This can be viewed as a cycle, whereby creative players perform well and make progress in the game, which enhances their affective state, leading them to be more creative.

In short, students with higher estimates of creativity perform better in the game, and they tend to enjoy the experience more than those with lower creativity estimates. But what about the effects of creativity on learning? We found that students with higher estimates of creativity tend to learn more content (physics) than those with lower creativity estimates. However, when controlling for incoming knowledge, creativity was not a significant predictor of posttest scores. Because creativity did predict game performance, and game performance predicts posttest scores (controlling for the pretest), we can say that creativity may exert an indirect effect on learning.[2] In other words, creativity, performance, and learning are connected. These relationships can be understood in the context of "flow." For example, Csikszentmihalyi (1997) has noted that when people experience the state of flow, they tend to be highly engaged and report that they enjoy the process while at the same time lose track of time. During the flow state, one can perform

[2] Although not reported in the paper, the results of a regression analysis showed that the number of levels completed (game performance variable) significantly predicted posttest scores controlling for pretest ($\beta = 0.16$, $t = 2.27$, $p = .03$; $F(2, 151) = 45.19$, $p < .001$, $R^2 = 0.37$) with an important effect of incoming physics knowledge on learning ($\beta = 0.54$, $t = 7.92$, $p < .001$).

at his or her best, and as a result learn the topic or skill at hand. Moreover, we know that affective states and creativity are strongly connected (Amabile, 1983). That is, when being creative (e.g., solving a level in *Physics Playground*) one's affective state impacts engagement, which in turn influences performance during gameplay, which in turn effects learning outcomes (Shute et al., 2015).

In this study, we decided to exclude behaviors that were considered as gaming-the-system behaviors, like the "stacking" solution shown in Fig. 3. We acknowledge that such solutions could be examined and analyzed related to creativity in future research. Specifically, it could be informative to investigate the frequency of such behaviors in gameplay and creativity. We hypothesize that once a player finds a way to game the system, they may stick to those behaviors and this may hinder their creativity as they would not think divergently when coming up with solutions. In fact, this hypothesis was one of the reasons we decided to exclude these behaviors from our analyses. Future research aiming to investigate the relationship between these behaviors and creativity can employ a pretest and posttest of creativity to see if the frequency of such behaviors impact creativity positively or negatively.

More work is needed in the area of automated assessment of creativity. Studies such as the current one are just the beginning of an important research stream which can involve several fields (e.g., computer science, learning sciences, psychometrics, and creativity studies). The studies that aim to create valid and reliable assessments of creativity can benefit research related to *enhancing* creativity. For example, in the future, we could have video games that use assessment methods such as stealth assessment to diagnostically assess and adaptively support a player's creativity. However, more research is needed that examines the specific effects of video games on creativity in general (e.g., meta-analyses), and of video games designed specifically to support creativity, like *Physics Playground* and other types of "sandbox" games. Such video games should be able to *facilitate* creative thinking processes while a person is actively engaged in constructing a creative solution. Towards that end, we just finished developing, and are currently testing a new creativity support system that resides in *Physics Playground's* level editor. The two support systems are based on two schools of thought—Inspirationalism and Structuralism (Shneiderman, 2009). Inspirationalists believe that during a task, creativity may be enhanced by getting inspired from seeing prior work in the area, using brainstorming strategies, making remote associations, using analogies, and other techniques intended to inspire one to be more creative. Structuralists believe that guiding people through a structured and orderly process can improve creativity. The current research is testing which of the two support types is more effective in enhancing students' creativity using four conditions: Inspiration support, Structural support, Combined Inspiration and Structural support, and a no-support control group.

In closing, games with effective creativity support systems can potentially serve as effective tools for enhancing creativity, particularly if coupled with valid real-time measures of creativity. Enhancing people's creativity via video games is a viable goal that creativity researchers should pursue—limited only by their imagination.

Declaration of interest

We wish to confirm that there are no known conflicts of interest associated with this manuscript.

CRediT authorship contribution statement

Valerie J. Shute: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Seyedahmad Rahimi:** Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing - original draft, Writing - review & editing.

Acknowledgments

This work was supported by the US National Science Foundation [award number #037988] and the US Department of Education [award number #039019]. We'd also like to thank team members who helped in this project—Russell Almond, Mathew Ventura, Yoon Jeon Kim, Lubin Wang, and Weinan Zhao.

Appendix A

Self-report Openness Scale.
(from 1 = strongly disagree to 5 = strongly agree)

1. I like to think of new ideas
2. I enjoy art
3. I am excited by many different activities
4. I daydream a lot
5. I enjoy learning new things
6. I like to explore different solutions to problems
7. I have an active imagination
8. I like to be original
9. I try to be different from other students
10. I am curious about many different things

References

Abdulla, A. M., & Cramond, B. (2017). After six decades of systematic study of creativity: What do teachers need to know about what it is and how it is measured? *Roeper Review, 39*(1), 9–23. https://doi.org/10.1080/02783193.2016.1247398

Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). *Improving stealth assessment in game-based learning with LSTM-based analytics*. International Educational Data Mining Society.

Almond, R. G. (2010). I can name that Bayesian network in two matrixes! *International Journal of Approximate Reasoning, 51*(2), 167–178.

Almond, R. G. (2015). An IRT-based parameterization for conditional probability tables. In J. M. Augusta, & R. N. Carvalho (Eds.), *Proceedings of the 2015 bayesian modeling application workshop at the 2015 uncertainty in artificial intelligence conference*. Retrieved from http://pluto.coe.fsu.edu/RNetica/.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology, 45*(2), 357. https://doi.org/10.1037/0022-3514.45.2.357

Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior, 36*, 157–183. https://doi.org/10.1016/j.riob.2016.10.001

Baer, J., & Kaufman, J. C. (2005). Bridging generality and specificity: The amusement park theoretical (APT) model of creativity. *Roeper Review, 27*(3), 158–163.

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning, 32*, 267–280.

Barojas, J., & Pérez, R. P. Y. (2001). Physics and creativity: Problem solving and learning contexts. *Industry and Higher Education, 15*(6), 431–439.

Berzonsky, M. D., & Sullivan, C. (1992). Social-cognitive aspects of identity style: Need for cognition, experiential openness, and introspection. *Journal of Adolescent Research, 7*(2), 140–155. https://doi.org/10.1177/074355489272002

Cheng, V. M. (2004). Developing physics learning activities for fostering student creativity in Hong Kong context. In *, Vol. 5. Asia-pacific forum on science learning and teaching* (pp. 1–33). The Education University of Hong Kong, Department of Science and Environmental Studies. No. 2.

Cheng, M.-T., Lin, Y.-W., & She, H.-C. (2015). Learning through playing virtual age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters. *Computers & Education, 86*(1), 18–29.

Cheung, P. C., Lau, S., Chan, D. W., & Wu, W. Y. (2004). Creative potential of school children in Hong Kong: Norms of the Wallach-Kogan Creativity Tests and their implications. *Creativity Research Journal, 16*(1), 69–78.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653–665. https://doi.org/10.1016/0191-8869(92)90236-I

Cropley, A. J. (1968). A note on the Wallach-Kogan tests of creativity. *British Journal of Educational Psychology, 38*(2), 197–201.

Cropley, A. J. (1972). Originality scores under timed and untimed conditions. *Australian Journal of Psychology, 24*(1), 31–36.

Cropley, D. H., & Cropley, A. J. (2011). Aesthetics and creativity. In M. A. Runco, & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (pp. 24–28). San Diego, CA: Academic Press/Elsevier.

Csikszentmihalyi, M. (1997). *Creativity: Flow and the psychology of discovery and invention*. New York: Basic Books.

DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction, 2012*.

Dewett, T. (2007). Linking intrinsic motivation, risk taking, and employee creativity in an R&D environment. *R & D Management, 37*(3), 197–208. https://doi.org/10.1111/j.1467-9310.2007.00469.x

Diakidoy, I. A. N., & Constantinou, C. P. (2001). Creativity in physics: Response fluency and task specificity. *Creativity Research Journal, 13*(3-4), 401–410.

Dippo, C. (2013). Evaluating the alternative uses test of creativity. In *Proceedings of the national conference, on undergraduate research (NCUR) 2013*. WI: University of Wisconsin La Crosse.

Entertainment Software Association. (2019). *Essential facts about the computer and video game industry*.

Feist, G. J. (1999). The influence of personality on artistic and scientific creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 273–296). Cambridge University Press.

Fessakis, G., & Lappas, D. (2013). Cultivating preschoolers creativity using guided interaction with problem solving computer games. In C. Carvallo, & P. Escudeiro (Eds.), *Vol. 2. Proceedings of the 7th European conference on games based learning (ECGBL2013)* (pp. 763–770).

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *2*(1), 5–16.

Glover, J. A. (1977). Risky shift and creativity. *Social Behavior and Personality: International Journal, 5*(2), 317–320. https://doi.org/10.2224/sbp.1977.5.2.317

Glover, J. A., & Sautter, F. (1977). Relation of four components of creativity to risk-taking preferences. *Psychological Reports, 41*(1), 227–230. https://doi.org/10.2466/pr0.1977.41.1.227

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *The Journal of the Learning Sciences, 22*, 521–563.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53*(4), 267. https://doi.org/10.1037/h0040755

Hamlen, K. R. (2009). Relationships between computer and video game play and creativity among upper elementary school students. *Journal of Educational Computing Research, 40*(1), 1–21. https://doi.org/10.2190/EC.40.1.a

Hewitt, R. (2012). *SpinCycle project [Weblog]*. Retrieved July 23, 2019, from SpinCycle project website: https://spincycleproject.wordpress.com/.

Inchamnan, W., Wyeth, P., & Johnson, D. (2013). Does activity in computer game play have an impact on creative behaviour? *2013*. In *IEEE international games innovation conference (IGIC)* (pp. 77–84). https://doi.org/10.1109/IGIC.2013.6659169

Jackson, P. W., & Messick, S. (1965). The person, the product, and the response: Conceptual problems in the assessment of creativity. *Journal of Personality, 33*(3), 309–329. https://doi.org/10.1111/j.1467-6494.1965.tb01389.x

Jackson, L. A., Witt, E. A., Games, A. I., Fitzgerald, H. E., von Eye, A., & Zhao, Y. (2012). Information technology use and creativity: Findings from the children and technology project. *Computers in Human Behavior, 28*(2), 370–376. https://doi.org/10.1016/j.chb.2011.10.006

Kanar, A. M., & Bell, B. S. (2013). Guiding learners through technology-based instruction: The effects of adaptive guidance design and individual differences on learning over time. *Journal of Educational Psychology, 105*, 1067–1108.

Kaufman, J. C., & Beghetto, R. A. (2013). Do people recognize the four Cs? Examining layperson conceptions of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 7*(3), 229.

Kaufman, J. C., & Sternberg, R. J. (2007). Resource review: Creativity. *Change, 39*(4), 55–58. Retrieved from JSTOR.

Ke, F., Parajuli, B., & Smith, D. (2019). Assessing game-based mathematics learning in action. In *Game-based assessment revisited* (pp. 213–227). Cham: Springer.

Ke, F., & Shute, V. (2015). Design of game-based stealth assessment and learning support. In *Serious games analytics* (pp. 301–318). Cham: Springer.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing, 16*(2), 142–163.

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a bayesian network example. *Computers in Education, 85*, 23–34.

Kloonigames. (2014). *Crayon physics Deluxe*. http://www.crayonphysics.com/.

Kogan, N., & Wallach, M. A. (1964). *Risk taking: A study in cognition and personality*. Oxford, England: Holt, Rinehart & Winston.

Kücklich, J. (2004). *Other playings: Cheating in computer games*.

Lubart, T. I. (1994). *Product-centered self-evaluation and the creative process*. Ph.D., Yale University. Retrieved from https://search.proquest.com/pqdtglobal/docview/304117293/abstract/5EF1668A69B2428EPQ/1.

MacCrimmon, K. R., & Wehrung, D. A. (1990). Characteristics of risk taking executives. *Management Science, 36*(4), 422–435. Retrieved from JSTOR.

Madjar, N., Greenberg, E., & Chen, Z. (2011). Factors for radical creativity, incremental creativity, and routine, noncreative performance. *Journal of Applied Psychology, 96*(4), 730–743. https://doi.org/10.1037/a0022416

McClelland, D. C. (1973). Testing for competence rather than for "intelligence. *American Psychologist, 28*(1), 1–14. https://doi.org/10.1037/h0034092

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*(6), 1258. https://doi.org/10.1037/0022-3514.52.6.1258

McCrae, R. R. (1996). Social consequences of experiential openness. *Psychological Bulletin, 120*(3), 323. https://doi.org/10.1037/0033-2909.120.3.323

McCrae, R. R., & Costa, P. T., Jr. (1985). Openness to experience. *Perspectives in personality, 1*, 145–172.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., et al. (2015, June). DeepStealth: Leveraging deep learning models for stealth assessment in game-based learning environments. In *International conference on artificial intelligence in education* (pp. 277–286). Cham: Springer.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Moffat, D. C., Crombie, W., & Shabalina, O. (2017). Some video games can increase the player's creativity. *International Journal of Game-Based Learning, 7*(2), 35–46. https://doi.org/10.4018/IJGBL.2017040103

Partnership for 21st Century Learning. (2019). *Framework for 21st-century learning*. Retrieved from http://www.nea.org/home/34888.htm.

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist, 39*(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman, & R. J. Sternberg (Eds.), *The cambridge handbook of creativity* (pp. 48–73). https://doi.org/10.1017/CBO9780511763205.005

Rahimi, S., & Shute, V. J. (in press). The effects of video games on creativity: A systematic review. In S. W. Russ, J. D. Hoffmann, & J. C. Kaufman (Ed.), Handbook of lifespan development of creativity (37 pages). Cambridge: Cambridge University Press.

Richards, R. (1990). Everyday creativity, eminent creativity, and health: "Afterview"; for CRJ issues on creativity and health. *Creativity Research Journal, 3*(4), 300–326. https://doi.org/10.1080/10400419009534363

Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2010). *Generation M 2: Media in the Lives of 8-to 18-Year-Olds*. Menlo Park, California: Henry J. Kaiser Family Foundation.

Runco, M. A. (2003). Education for creative potential. *Scandinavian Journal of Educational Research, 47*(3), 317–324.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75.

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education, 69*, 452–462.

Sawyer, R. K. (2011). *Explaining creativity: The science of human innovation*. New York, USA: Oxford university press.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age*. MIT Press.

Shneiderman, B. (2009). Creativity support tools: A grand challenge for HCI researchers. In M. Redondo, C. Bravo, & M. Ortega (Eds.), *Engineering the user interface* (pp. 1–9). https://doi.org/10.1007/978-1-84800-136-7_1

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., Rahimi S., Smith, G., Ke, F., Almond, R., Dai, C-P, Kamikabeya, R., Liu, Z., Yang, X., & Sun, C. (in press). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in Physics Playground. Journal of Computer-Assisted Learning (38 pages).

Shute, V. J., D'Mello, S. K., Baker, R., Bosch, N., Ocumpaugh, J., Ventura, M., et al. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224–235.

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

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. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321).

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. https://doi.org/10.1080/00220671.2013.832970

Shute, V. J., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs. In A. A. Rupp, & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and application* (pp. 535–562). Hoboken, NJ: John Wiley & Sons, Inc.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. R. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117.

Silvia, P. J. (2008a). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts, 2*(3), 139–146. https://doi.org/10.1037/1931-3896.2.3.139

Silvia, P. J. (2008b). Creativity and intelligence revisited: A latent variable analysis of Wallach and kogan. *Creativity Research Journal, 20*(1), 34–39. https://doi.org/10.1080/10400410701841807

Sternberg, R. J., & Lubart, T. I. (1992). Buy low and sell high: An investment approach to creativity. *Current Directions in Psychological Science, 1*(1), 1–5.

Sternberg, R. J., & Lubart, T. I. (1996). Investing in creativity. *American Psychologist, 51*(7), 677. https://doi.org/10.1037/0003-066X.51.7.677

The State of Online Gaming. (2019). *Market research: The state of online gaming - 2019*. https://www.limelight.com/resources/white-paper/state-of-online-gaming-2019.

Tingir, S. & Almond, R. (in press). An augmented EM algorithm for monotonic Bayesian networks using parameterized conditional probability tables. To appear in *Behaviormetrika* (33 pages).

Torrance, E. P. (1972). Predictive validity of the Torrance tests of creative thinking. *Journal of Creative Behavior, 6*(4), 236–262. https://doi.org/10.1002/j.2162-6057.1972.tb00936.x

Torrance, E. P. (1993). Understanding creativity: Where to start? *Psychological Inquiry, 4*(3), 232–234. https://doi.org/10.1207/s15327965pli0403_17

Velikovsky, J. T. (2014). Flow theory, evolution & creativity: Or, 'fun & games. In *Proceedings of the 2014 conference on interactive entertainment* (pp. 1–10).

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children.* New York, NY: Holt, Rinehart & Winston.

Wijman, T. (2019). *The global Games Market will generate $152.1 billion in 2019 as the U.S. Overtakes China as the Biggest Market.* https://newzoo.com/insights/articles/the-global-games-market-will-generate-152-1-billion-in-2019-as-the-u-s-overtakes-china-as-the-biggest-market/.

Wright, W. (2006, April 1). Dream machines. *Wired, 14*(4), 110–112.