22

Assessing and Supporting Hard-to-Measure Constructs in Video Games

Valerie Shute and Lubin Wang

UNCORRECTED PROOFS

Imagine being asked to design an assessment as part of a class assignment. You have a choice of topics to measure: math skills (e.g., two-digit subtraction) or creativity (e.g., relative to creative problem solving). To get an "A" on the assignment, your assessment must be valid and reliable. Most of us would choose to develop the math assessment because designing a creativity assessment is really hard. Or is it?

Consider the image in Figure 22.1 that recently made its way through the Internet. The first two panels comprise the original image showing graphical representations for "knowledge" and "experience." A week or two after the image initially appeared, a person tacked on a third panel to the pair, labeling it "creativity."

The majority of us would agree that the newly altered image clearly provides positive evidence of creativity, and is likely more valid than evidence coming from responses to a series of self-report questions addressing creativity (e.g., "I like to think of new ideas"). To understand the systemic implications of this example more fully we briefly discuss the connections between constructs, assessment, and learning illustrated by this figure.

Constructs

In education and psychology, the term **construct** typically refers to a complex psychological concept such as mathematics skills, reading ability, visual-spatial processing, collaboration, curiosity, intelligence, creativity, conscientiousness, happiness, and anxiety. Generally speaking, we cannot directly measure constructs like we measure our weight and height because constructs are theoretical conceptualizations. However, we can make meaningful indirect inferences about an individual's level of a particular construct based on an accumulation of targeted things they say and do in relevant circumstances (i.e., response data that contain evidence about constructs).

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.
© 2016 John Wiley & Sons, Inc. Published 2016 by John Wiley & Sons, Inc.

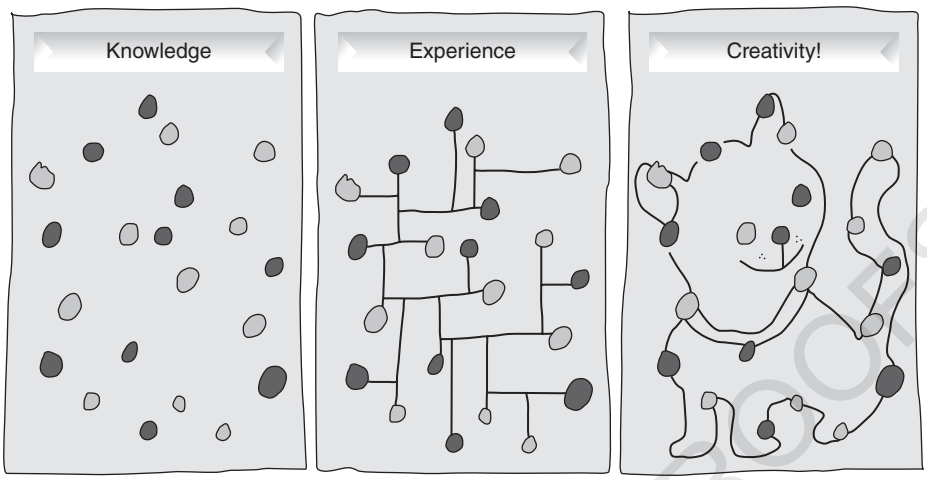

Figure 22.1 Viral Internet image, circa 2014.

Some constructs such as creativity have historically been deemed "hard to measure" for a variety of reasons (e.g., lack of a clear and agreed-upon definition, psychological and/or statistical multidimensionality of the construct, subjectivity of scoring, and so on). A few examples of **hard-to-measure constructs** that we have assessed in our own work include creativity (see Kim & Shute, 2015), problem solving (see Shute, Ventura, & Ke, 2015), persistence (see Ventura & Shute, 2013), systems thinking (see Shute, Masduki, & Donmez, 2010), **gaming-the-system** (see Baker, 2005; Wang, Kim, & Shute, 2013), and design thinking (see Razzouk & Shute, 2012), among others. Our general focus in this chapter is on optimal ways to assess hard-to-measure constructs, specifically within the context of well-designed digital games. Our premise is that good games, coupled with evidence-based embedded assessment, show promise as a means to dynamically assess hard-to-measure constructs more accurately and decidedly more engagingly than traditional approaches.

Assessment

Assessment involves more than just measurement. The term refers to collecting, analyzing, and interpreting information about a student's understanding and/or performance relative to educational goals. An assessment can use a variety of procedures for evaluating student work and learning and thus represents a general collection of tools that includes standardized tests. For example, if one is interested in determining a student's progress toward educational goals one can: (a) administer a test, (b) view a collection of relevant student work, (c) ask the student to evaluate her progress, (d) observe the student solve a complex task with manipulables, (e) examine and evaluate **log file** data from a digital environment, and so on.

That is, assessment is both an instrument and a process by which information is obtained relative to a known objective. But because inferences are made about what a person knows on the basis of her responses to a finite number of assessment tasks,

UNCORRECTED PROOFS

there is always some uncertainty in inferences made on the basis of assessments. The goal in educational measurement is to collect relevant information about a student or a collection of students and to minimize uncertainty or error variance around reported information such as scores while doing so.

Consequently, key aspects of assessment quality are **reliability** and **validity**. *Reliability* refers to the consistency of assessment results across conditions like alternative assessment forms, testing occasions, or testing environments. *Validity* refers to the extent to which the assessment accurately measures what it is supposed to measure, and the accuracy and defensibility of the inferences made from task or test results.

There are three closely related educational and political functions of assessment. The first function involves "closing the loop," or using actionable results to improve learning and make good decisions about what to do next – by the student, teacher, administrator, or other stakeholders. This is true whether the purpose of the assessment is to support student learning (formative uses) or accountability (summative uses).

A second related function of assessment is to "make student learning visible." As we mentioned earlier, you cannot assess what another person knows, can do, feels, or believes unless there is some observable evidence of that learning. Establishing how to make learning visible, however, is difficult. Most of a person's knowledge (and other mental states and traits) is invisible to others, and sometimes even to oneself. Because a person's thoughts cannot be seen, one depends on indicators that suggest the nature or status of his or her knowledge.

A third important function of assessment is "diagnostic." That is, assessment is part of a process used to determine students' strengths and weaknesses in relation to educationally valuable competencies. As such, assessment provides a way to figure out the nature and extent of difficulties in a student's understanding or problem-solving efforts. To begin to use assessment to support student learning, we need to design tasks so that this information can be disentangled and interpreted in valid and reliable ways (see Hunt & Minstrell, 1996; Minstrell, 2001, for more on this topic).

In line with sound assessment design principles, a good diagnostic assessment system should allow the user to be able to infer competency estimates accurately for a student at various targeted "**grain sizes**" (i.e., the scope or generality of the competencies; see McCalla & Greer, 1994) to serve these three functions. This process begins with the design of a reasonable (i.e., accurate and informative) competency model, which provides the basis for task-level (i.e., real-time, formative) and overall (i.e., summative) diagnoses to occur (e.g., Jang, 2009; Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; Shute, Ventura, Bauer, & Zapata-Rivera, 2009; VanLehn, 2006).

Task-level diagnoses can provide immediate support to the student via task-specific feedback. Estimates of more general competencies provide the basis for decisions concerning what to do next such as selecting a new task or offering other content to the student, providing practice, or some other instructionally helpful activities. This can be accomplished automatically behind the scenes via computational selection rules and algorithms. Alternatively, diagnostic results can be provided to the teacher in the form of instructional prescriptions or suggestions about what to do next (e.g., proceed to the next topic or spend more time helping students understand the current topic).

UNCORRECTED PROOFS

Learning

Learning refers to the acquisition of new knowledge and skills as well as other personal **attributes**, and is generally regarded as a *constructive activity*. However, the construction can assume many forms. Individuals differ in how they learn (processes) as well as what they learn (outcomes). In addition, there are many other variables that affect both processes and outcomes including abilities, interests, backgrounds, prior knowledge, skills, personality traits, affective states, self-efficacy, and motivation (e.g., Shute, Lajoie, & Gluck, 2000).

Advances in the learning sciences suggest that acquiring and demonstrating new knowledge and skills occurs within an environment or context, which includes learners with specific cognitive and affective profiles and tools to support and assess learning (Pellegrino, Chudowsky, & Glaser, 2001). Assessment can play a key role in facilitating learning. That is, when we assess student learning, we are asking the question, "What have the students learned and how well have they learned it?"

Organization of Chapter

The rest of the chapter is split into six sections as follows. In the next section we identify some specific hard-to-measure constructs and describe what makes them especially difficult to measure. In the second section we argue for the use of well-designed games to assess and support such hard-to-measure constructs. In the third section we recommend a particular assessment design framework called **evidence-centered design** (ECD; see Mislevy, Steinberg, & Almond, 2003) coupled with a form of embedded assessment called **stealth assessment** (e.g., Shute, 2011) to measure and support learning in games. In the fourth section, we illustrate two projects that involve the stealth assessment of two hard-to-measure constructs (problem solving and creativity) within two different games (*Plants vs. Zombies 2* and *Physics Playground*). The fifth section focuses on psychometric rigor, especially relative to establishing the reliability and validity of stealth assessments. Finally, we conclude with a brief discussion of the obstacles to surmount in this line of research and the future research needed.

Hard-to-measure Constructs and Why They are Hard to Measure

Hard-to-Measure Constructs

With the emergence of the Internet circa mid-1990s, the world became more interconnected, effectively smaller, and more complex than before (Friedman, 2005). Developed countries now rely on their knowledge workers to deal with an array of complex problems, many with global ramifications (e.g., climate change or renewable energy sources). When confronted by such problems, tomorrow's workers need to be able to think systemically, creatively, and critically (see, e.g., Shute & Torres, 2012; Walberg & Stariha, 1992). These skills are a few of what many

UNCORRECTED PROOFS

educators are calling **twenty-first-century (or complex) competencies** (see Partnership for 21st Century Learning, 2012; Trilling & Fadel, 2009).

Ensuring that students at any age succeed in the twenty-first century requires fresh thinking about what knowledge, skills, and other personal attributes (i.e., what we call, collectively, competencies) are or should be supported in our schools. In addition, there is a need to design and develop valid assessments to measure and support these competencies. Except in rare instances, our current education system neither teaches nor assesses these new competencies despite a growing body of research showing that competencies such as persistence, creativity, self-efficacy, openness, and teamwork can substantially impact student academic achievement (Noftle & Robins, 2007; O'Connor & Paunonen, 2007; Poropat, 2009; Sternberg, 2006; Trapmann, Hell, Hirn, & Schuler, 2007). As a result, many of our current assessments fail to assess what students actually can do with the school-acquired knowledge and skills (Shute, 2009).

Measurement Challenges

Key measurement challenges for hard-to-measure constructs include (a) lack of a clear and consensual definition and/or **operationalization** of the construct, (b) theoretical multidimensionality of the construct where certain dimensions may have internal as well as external sources (e.g., see Jirout & Klahr, 2012 for problems relating to measuring curiosity), (c) difficulty disambiguating trait from state (e.g., anxiety and creativity – where some people tend to be generally anxious and/or creative, and others are only anxious and/or creative in certain settings or domains), (d) difficulty disambiguating the generality of the construct (e.g., is there a single "persistence" variable, or is persistence solely dependent on the context), and (e) reliance on outdated multiple-choice and self-report measures, the former narrowly focused and the latter flawed.

Self-report measures in particular are subject to "social desirability effects" that can lead to false reports about behaviors, attitudes, and beliefs (see Paulhus, 1991). In addition, test takers may interpret specific self-report **items** differently (e.g., what it means to "work hard") leading to unreliability and lower validity (Lanyon & Goodstein, 1997). Finally, self-report items often require that individuals have explicit knowledge of their dispositions (see, e.g., Schmitt, 1994), which is not always the case.

What we need are new valid **performance-based assessments** that assess how students use complex competencies that are directly relevant for the real world. One challenge with developing a performance-based assessment for a hard-to-measure construct is crafting appropriate situations or problems to elicit the competency of interest. One way to approach this problem is to use **digital learning environments**, including video games, to simulate a variety of problems for performance-based assessment (Dede, 2005; DiCerbo & Behrens, 2012; Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Quellmalz, Timms, Silberglitt, & Buckley, 2012; Shute, 2009). Digital learning environments can provide meaningful assessment environments by supplying students with scenarios that require the application of various competencies.

UNCORRECTED PROOFS

Well-designed Games as Vehicles for Assessing and Supporting These Constructs

According to the findings from a three-year ethnographic study on middle- and high-school students, playing video games with friends and family has become a daily routine (Ito et al., 2010) and using video games for educational purposes is becoming more common. While navigating game environments, players engage in higher-order thinking skills such as reasoning, decision making, and problem solving. Players often receive ongoing feedback in the form of scores or their in-game characters' progress along the storyline. Playing games is basically the process of developing and honing particular competencies required to advance in the game.

Moreover, players are often faced with various tasks that are challenging but ultimately attainable. Even when they get stuck, they experience what is called "pleasant frustration" (Gee, 2007) because that is the risk they chose to take, and solving very difficult problems yields a large sense of achievement. Interested educators and practitioners are beginning to recognize that certain video games can be great educational tools; however, not so many realize yet the potential of games as an assessment vehicle.

Well-designed games can provide meaningful assessment environments by providing students with scenarios or tasks that require the application of various competencies (e.g., creativity, problem solving, and persistence). Furthermore, there is a convergence between the core elements of a good game and the characteristics of productive learning (Shute, Rieber, & Van Eck, 2011). That is, learning is at its best when it is active, goal-oriented, contextualized, and interesting (e.g., Bransford, Brown, Cocking, Donovan, & Pellegrino, 2000; Bruner, 1961); thus, learning environments should be interactive, provide ongoing feedback, grab and hold attention, and have appropriate and adaptive levels of challenge – all features of good games.

Gee (2003) has argued that the secret of a good game is not its 3D graphics and other bells and whistles, but its underlying architecture in which each level dances around the outer limits of the player's abilities, seeking at every point to be hard enough to be just doable (see also Csikszentmihályi, 1990, on **flow** theory). Along the same line, psychologists (e.g., Vygotsky, 1978) have long argued that the best instruction hovers at the boundary of a student's competence. Moreover, a good game reinforces a sense of control – a critical metacognitive component for self-regulated learning (Zimmerman & Schunk, 2001). Finally, both well-designed games and productive learning processes employ ongoing feedback as a major mechanism of play/learning support. All of these features of well-designed games warrant them as an appropriate vehicle to assess and support hard-to-measure constructs.

Consider the hard-to-measure construct of persistence as an illustration. Persistence can be broadly defined as the motivation to work hard despite challenging conditions and is considered to be a facet of conscientiousness. Conscientiousness has consistently been found to predict academic achievement from preschool (Abe, 2005) to high school (e.g., Poropat, 2009) to the postsecondary level (e.g., Noftle & Robins, 2007) and adulthood (e.g., Shiner, Masten, & Roberts, 2003). The traditional way to measure persistence is via self-report which, as discussed above, has limitations. However, new performance-based

UNCORRECTED PROOFS

methodologies record and score actual behaviors pertaining to that particular competency within an interactive context like a game.

In *Physics Playground* (formerly "Newton's Playground," see Shute & Ventura, 2013), actions that a player takes that inform the persistence variable include (a) how long a person spends on a difficult problem (where longer = more persistent), (b) the number of failures and re-tries before success, (c) incidences of returning to a hard problem after skipping it, and so on. Each instance of these "persistence indicators" updates the **student model** of this variable. Similarly, indicators for other facets of conscientiousness (e.g., perfectionism and organization) are captured and in turn update the overall model of conscientiousness. Over time, with a sufficient amount of **evidence accumulated**, the psychometric properties of the assessment (especially relative to reliability and validity) increase.

Evidence-centered Design and Stealth Assessment in Well-designed Games

As we discussed in the previous section, one main challenge for educators who want to employ or design games to support learning is making valid inferences – about what the student knows, believes, and can do – at any time, at various levels, and without disrupting the flow of the game (and hence engagement and learning). One way to increase the quality and utility of an assessment is to use ECD which informs the design of valid assessments and yields real-time estimates of students' competency levels across a range of knowledge and skills (Mislevy et al., 2003).

Evidence-Centered Design (ECD)

The ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about personal competencies via an evidentiary chain (i.e., valid arguments that serve to connect task performance to competency estimates) and are thus valid for their intended purposes. ECD consists of several conceptual and computational models that work in concert where the goal is to help assessment designers coherently align (a) the claims that they want to make about learners and (b) the things that learners say or do in relation to the contexts and tasks of interest (e.g., Mislevy & Haertel, 2006; Mislevy et al., 2003, and for a simple overview, see Shute, Kim, & Razzouk, 2010).

In ECD, after conducting a thorough domain analysis relative to the construct in question, the **conceptual assessment framework** (CAF; see Mislevy et al., 2003) is developed, consisting of five interrelated conceptual models. All five models are relevant to answering the questions of what-when-how to measure, but three of the five models are fundamental for structuring, operationalizing, planning, designing, and defending test-score inferences. The three main models in the ECD framework include the **competence/student model**, **evidence model**, and **task models**, described in more detail later. Once the CAF is formulated, it feeds into the **four-process architecture**

UNCORRECTED PROOFS

(see Almond, Steinberg, & Mislevy, 2002) which is a practical component in ECD for organizing and implementing the delivery of assessments.

Competency model (CM). What collection of knowledge, skills, and other attributes should be assessed? Variables in the *competency model* (CM) describe the set of knowledge and skills on which inferences are based (see Almond & Mislevy, 1999). Although ECD can work with simple one-dimensional competency models, its strength comes from treating a competency as multidimensional. The term *student model* (which comes from the **intelligent tutoring system** literature, see Shute & Psotka, 1996) is used to denote an 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 the level on each variable within the CM for a particular student.

Evidence model (EM): What behaviors or performances should reveal those competencies? 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 answer two questions, namely (a) what behaviors or performances reveal targeted competencies and (b) what is the statistical connection between those behaviors and the variables making up the CM? The EM consists of two main parts: (a) the **evidence rules** (sometimes called the **scoring model** or rubrics) and (b) the **statistical model** (sometimes called the **measurement model**).

The evidence rules take as input the **work product** resulting from the student's interaction with a task, which might be the selection of an answer option, a short answer, a graphical response, a series of actions taken to solve a game level, and so on. As output, the evidence rules produce **observable variables** (i.e., scores/indicators) that are evaluative summaries of the work products (e.g., "If the student has selected option D, then the score takes the value of correct; otherwise the score takes the value of incorrect"). In this case, the evidence rule takes the student's response, or "work product," to produce a score of correct or incorrect. Scoring can handle more complex cases with more than just dichotomous options of correct/incorrect.

Second, the statistical model expresses the relationship, in probability or logic, between the competency model variables and the observable variables (scores). It enables updating the competency model variables in a way that aggregates scores across tasks or performances. The statistical model may be as simple as number-right scoring for a single competency variable or may employ multidimensional statistical approaches such as **Bayesian inference networks/Bayes nets** to update an overall competency and other competency variables (see Shute & Ventura, 2013 for an applied example).

Task model (TM): What tasks or problems should elicit those behaviors that comprise the evidence? The *task model* (TM) variables describe features of situations that will be used to elicit performance. A TM provides a framework for characterizing or constructing situations with which a student will interact to provide evidence about targeted aspects of competencies. The main purpose of tasks or problems is to elicit observable evidence about unobservable competencies. The EM serves as the glue between the TM and CM.

There are two main reasons why we believe that the ECD framework fits well with the assessment of learning in digital games. First, in games people learn in action

UNCORRECTED PROOFS

(Gee, 2003; Salen & Zimmerman, 2005). That is, learning involves continuous interactions between the learner and the game, so learning is inherently situated in context. Therefore, the interpretation of knowledge and skills as the products of learning cannot be isolated from the context, and neither should assessment. The ECD framework helps us to link what we want to assess to what learners do in complex contexts. Consequently, an assessment can be clearly tied to learners' actions within digital games, and can operate without interrupting what they are doing or thinking (Shute, 2011).

The second reason that ECD is believed to work well with digital games is because it is based on the assumption that assessment is, at its core, an evidentiary argument. Its strength resides in the development of performance-based assessments where what is being assessed is latent or not apparent (Rupp, Gushta, Mislevy, & Shaffer, 2010). In many cases, it is not clear what people learn in digital games generally. However, with the help of ECD, we can figure out just what we want to assess (i.e., the claims we want to make about learners) and clarify the intended goals, processes, and outcomes of learning so that suitable stealth assessment processes are created that support educational objectives in a meaningful and data-driven way.

Stealth Assessment

For many people, tests are a source of anxiety. Test anxiety can have adverse effects on performance. New directions in educational and psychological measurement allow more accurate estimations of students' competencies, and new technologies let us administer formative assessments during the learning process, extract ongoing, multifaceted information from a learner, and react in immediate and helpful ways, as needed. As mentioned earlier, when embedded assessments are directly woven into the fabric of the learning or gaming environment so that they are virtually invisible, we call that *stealth assessment* (e.g., Shute et al., 2009).

The process of stealth assessment ensures that the assessment will neither disrupt flow (Csikszentmihályi, 1990) as learners fully engage in the gaming environment nor that it will cause anxiety that is often associated with traditional types of assessment. Such assessments are intended to support learning and remove (or seriously reduce) test anxiety while not sacrificing validity and reliability (e.g., DiCerbo & Behrens, 2012; Shute, Hansen, & Almond, 2008). Moreover, stealth assessments are supported by **automated scoring** and machine-learning techniques to infer things that would be too hard for humans (e.g., estimating values of competencies across a network of skills).

Stealth Assessment and ECD

In digital gaming environments with stealth assessment, the CM accumulates and represents belief about the targeted aspects of competencies, expressed as probability distributions for CM variables (Almond & Mislevy, 1999). EMs identify what the student says or does that can provide evidence about those skills (Steinberg & Gitomer, 1996) and express in a psychometric model how the evidence depends on the CM variables (Mislevy, 1994). TMs express situations that can evoke required evidence.

UNCORRECTED PROOFS

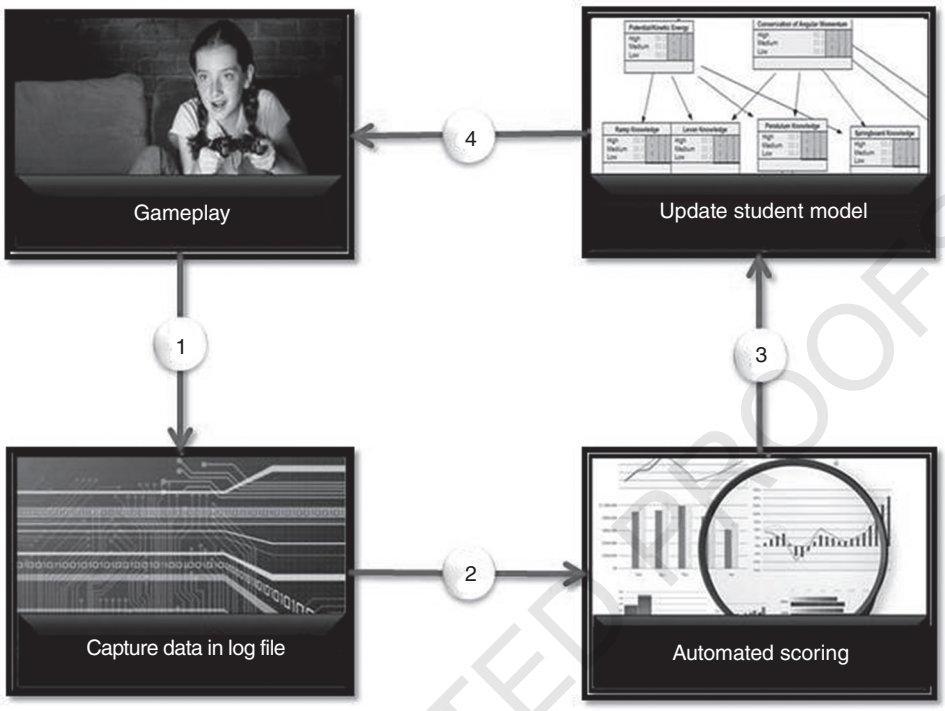

Figure 22.2 Stealth assessment cycle.

One big question is not about how to collect this rich digital data stream, but rather how to make sense of what can potentially become a deluge of information.

As shown in Figure 22.2, as students interact with tasks/problems in a game during the solution process they are providing a continuous stream of data captured in a log file (arrow 1) that is analyzed by the evidence model (arrow 2). The results of this analysis are data (e.g., scores or classifications) that are passed to the competency model, which statistically updates the claims about relevant competencies in the student model (arrow 3).

The estimates of competency levels in CM variables can also be used diagnostically and formatively to provide feedback and other forms of learning support to students as they continue to engage in gameplay (arrow 4). This process of making valid inferences about competency states and then using that information as the basis for offering learning support to the student is important to support the growth of constructs.

Returning to our original premise, we believe that well-designed games with evidence-based stealth assessment can be used to assess hard-to-measure constructs in real time, more accurately and engagingly than traditional approaches (see Shute, Ventura, & Kim, 2013). Additionally, we posit that good games provide an environment that can potentially improve various competencies including hard-to-measure constructs such as persistence. For example, games contain many problems that require players to persevere despite failure and frustration. That is, many good games can be quite difficult, and pushing one's limits is an excellent way to improve

UNCORRECTED PROOFS

persistence, especially when accompanied by the great sense of satisfaction one gets on successful completion of a thorny problem (see, e.g., Eisenberger, 1992; Eisenberger & Leonard, 1980).

Examples of Stealth Assessments of Hard-to-Measure Constructs in Two Games

To garner more acceptance of the idea of using games as assessments – especially for hard-to-measure constructs – we having been designing, developing, validating, and testing various stealth assessments for the past eight or so years. Our focus in this section is to introduce two examples of using video games as stealth assessment vehicles to measure two hard-to-measure constructs – problem-solving skills and creativity.

Stealth Assessment of Problem Solving Skills in Plants vs. Zombies 2

As Prensky (2005) has argued, video games provide a meaningful context for problem solving where players learn the rules of a particular game – what to do and what not to do in order to solve problems. Thus games can be a natural medium to assess problem solving skills. We recently worked with the GlassLab (https://www.glasslabgames.org/) on a project to design stealth assessments of problem-solving skills in the game *Plants vs. Zombies 2* (PvZ2). PvZ2 was developed by PopCap Games and published by Electronic Arts (July 2013) on the heels of the very successful first game of the series, Plants vs. Zombies. PvZ2 is a tower defense game utilizing a time-travel theme wherein players are invited to a number of different worlds (e.g., Ancient Egypt, Pirate Seas, Wild West, and Far Future) across different eras with around 25 challenging levels per world.

The goal of this game is to fight off approaching zombies by growing powerful plants in the limited soil squares in front of the home base. Each world contains a standard set of zombies along with some new zombie types that are tougher than the original ones and that possess special abilities. Players are also offered new types of plants to place in the given space in front of their home base. Different types of plants have different powers and so they should be chosen and planted wisely to defeat zombies effectively and efficiently. Plants can gain additional power for a short period of time after a player drag-and-drops plant food onto them. New types of zombies and plants are acquired as the player advances through the game; see Figure 22.3 for a screenshot of the game.

We created a problem-solving competency model after conducting a thorough review of the relevant literature (see Shute & Wang, 2015) and observing a number of experts' gameplay solutions on YouTube; we also aligned the competency model variables to the *Common Core State Standards*. That is, we selected standards (for grades 6–8) that are highly related to problem-solving skill for inclusion in our problem-solving competency model.

We ended up with four primary problem-solving facets, (1) understanding the givens and constraints in a problem, (2) planning a solution pathway, (3) using tools effectively/efficiently during solution attempts, and (4) monitoring and evaluating progress. Our CM for problem-solving within games of this type of genre (i.e., tower

UNCORRECTED PROOFS

Figure 22.3 Screen capture of PvZ2.

defense, strategy games) is thus conceptualized as a combination of the four main variables along with their relevant indicators (i.e., what the player does in the game that provides evidence for each facet). After finalizing the CM, we delineated connections among the problem solving facets (theoretical) and the gameplay activities (observable indicators) for establishing the EM, which specifies relationships among the observable behaviors in the game and levels of the variables in the competency model.

Figure 22.4 displays an example of how we linked the CM variables with sample evidence (i.e., indicators) derived from gameplay. As players interact with the game, estimates related to their problem-solving skills are updated as ongoing evidence accrues. For example, players are expected to collect sun power to grow certain types of plants. Different plants cost different amounts of sun power to grow, usually a function of the plant's strength. Generally, at the beginning of the game, players are provided with a very limited amount of sun power. Players should use the initial sun power to grow sunflowers that will generate more sun power.

The lack of sun power at the beginning of the level is considered a constraint. If a player is not able to identify the constraint and plants something other than sunflowers early in the game, then the player would be scored "low" on this constraint facet and will likely fail that level because he or she will not have enough offensive plants to fight the approaching zombies later. Players who understand the constraint would likely plant as many sunflowers as quickly as possible at the beginning to allow for sustainable development.

UNCORRECTED PROOFS

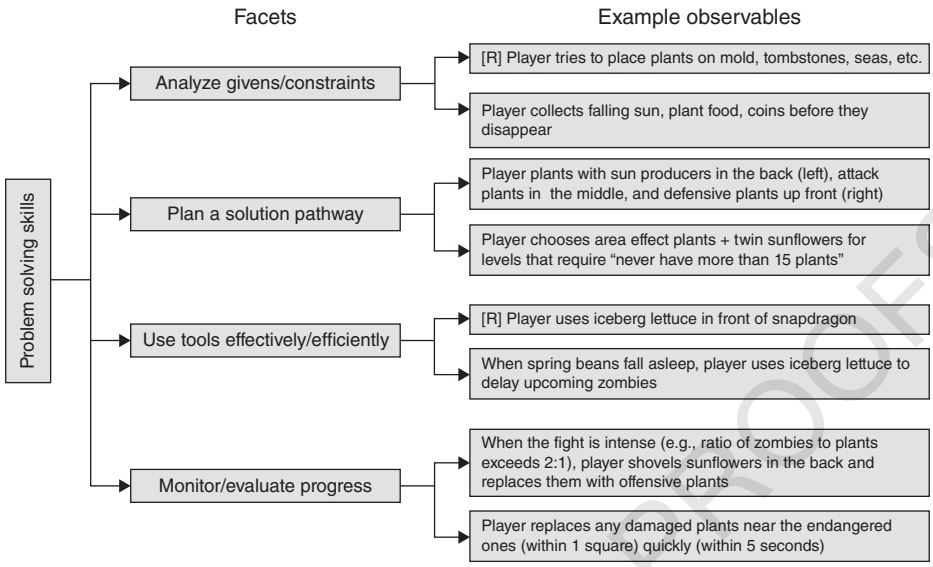

Figure 22.4 CM for problem solving in PvZ2; [R] means reverse coded.

To accumulate evidence across levels in the game, we use *Bayesian networks* or *Bayes nets* for short, which are graphical representations of the conditional dependencies between different variables (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015); we created our Bayes nets using the software Netica (by Norsys Software Corporation). In our Bayes nets for the level relating to the sunflowers example, the relevant indicator was defined as "Player plants at least three sunflowers at the beginning before the second wave of zombies shows up." There are two possible states of this indicator: "yes" (the player has accomplished this indicator in the current level) and "no" (the player did not accomplish the indicator). Successful completion of the indicator leads to an increase in the player's score on the specific skill "analyze givens and constraints" and on the "overall problem solving skill."

This is indicator no. 8 shown with a border in the lower-left side of Figure 22.5, which illustrates the full problem-solving Bayes net for level 9 of PvZ2. The upper five nodes represent the overall problem-solving estimate along with the four main skills or facets of interest, which are the theoretical constructs on which we are making inferences. The lower nodes represent the gameplay indicators, which are the observables that will be used as the basis for rendering inferences.

Table 22.1 shows the **conditional probability table** for indicator no. 8 and this skill variable in level 9 of "Ancient Egypt." We determined the distributions of **conditional probabilities** based on the DiBello-Samejima models, which includes the **item response theory** parameters of discrimination and difficulty (see Almond, 2010; Almond et al., 2015).

The discrimination estimate for this indicator/item was set to 0.1 (i.e., low) and the difficulty was set to −0.3 (i.e., fairly easy). These parameters were determined via data we collected from a pilot study using the game.

UNCORRECTED PROOFS

0002738796.INDD 548

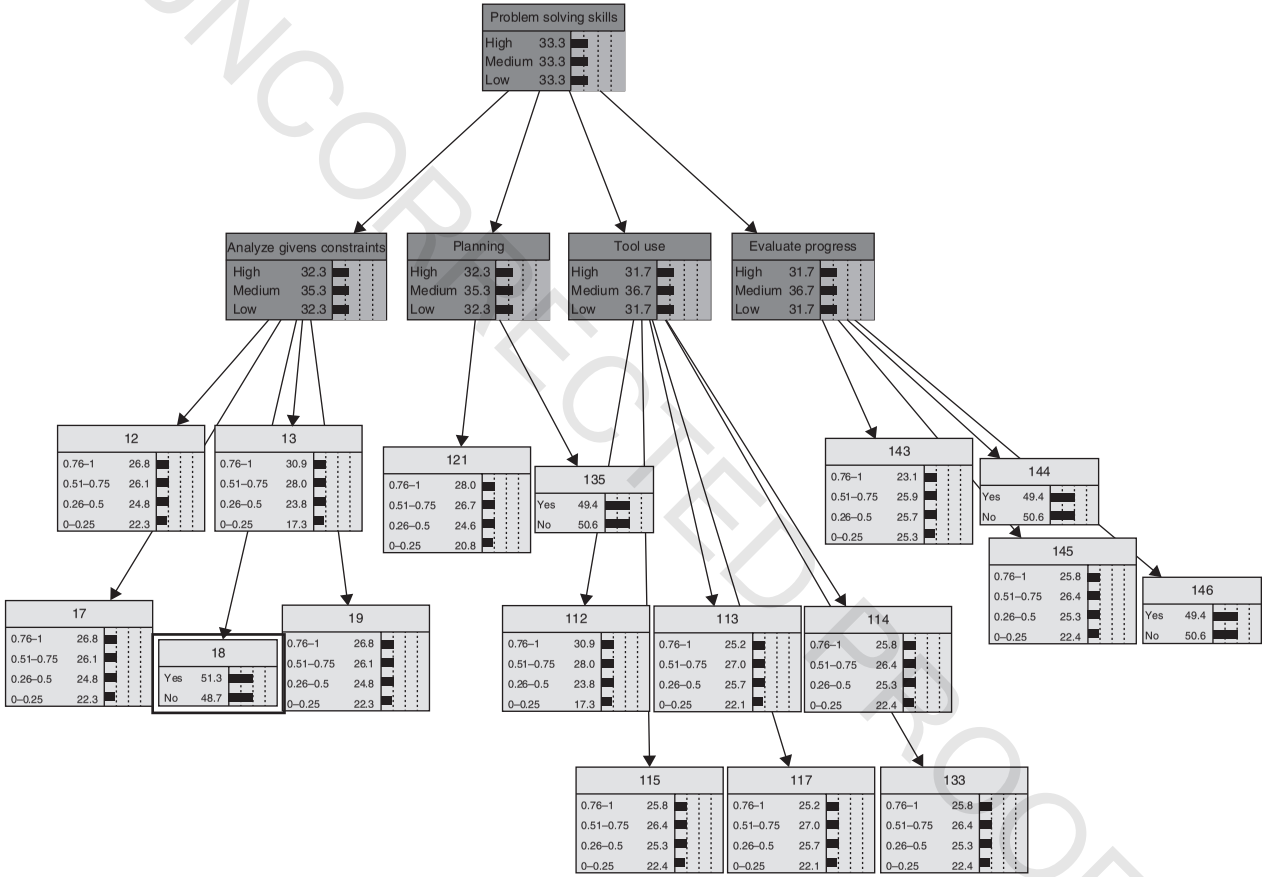

Figure 22.5 Initial Bayes net model with prior probabilities of problem solving for level 9 in PvZ2.

UNCORRECTED PROOFS

5/12/2016 4:00:00 PM

Table 22.1 Conditional probability table for indicator no. 8 and specific skill in Level 9.

Analyze givens and constraints	*Yes*	*No*
High	55.4	44.6
Medium	51.3	48.7
Low	47.2	52.8

All discrimination and difficulty parameters for each indicator in each level were specified in an *augmented* **Q-matrix** (Almond, 2010). In the basic format of the Q-matrix, the rows represent the indicators relevant for problem solving in each level and the columns represent the four key skills or facets of problem solving. If an indicator is relevant to a skill, the value of the cell is "1"; otherwise it is "0." We recommend always conducting a detailed task analysis before making the connection between competency variables and evidence so that proper indicators from the game can be identified to elicit specific evidence about the hard-to-measure construct. In the augmented version of the Q-matrix, we additionally specified the values of the difficulty and discrimination parameters of each indicator at different levels. The values of difficulty and discrimination were used to determine the conditional probabilities between the indicator and the facet, which is important for future reference (e.g., examination or adjustment of the values).

Data from the PvZ2 project are currently being analyzed. We completed a small pilot test and adjusted conditional probability tables in our original PvZ2 Bayes nets based on participants' performance data in the game as well as on an external measure of problem-solving skill the players used before gameplay called *MicroDYN* (Wustenberg, Greiff, & Funke, 2012). Once the game and its updated data-collection and accumulation processes are completed, our next step will be to conduct a full validation study of the stealth assessment problem solving measures against existing validated measures.

Specifically, we plan to use *Raven's progressive matrices* (Raven, 1962, 2000) and MicroDYN as our external measures of problem-solving skill and will correlate our in-game estimates of each of the four facets (as well as the overall measure of problem solving) against the external measures to test construct validity. Results from a small validation study ($n = 52$) show that our stealth assessment measure of problem-solving skills significantly correlates with both Raven's ($r = .40$; $p < .01$) and MicroDYN ($r = .48$; $p < .01$) suggesting convergent validity (for more details see Shute, Ke, and Wang, in press).

Stealth Assessment of Creativity in Physics Playground

The second hard-to-measure construct we elected to discuss in this chapter is creativity. Lubart (1994) defines creativity as the ability to produce work that is both novel and relevant. Creativity is identified as one of the most essential **twenty-first-century skills** that separate students who are prepared for complex and challenging life and work environments from those who are not (Partnership for 21st Century Learning, 2012). The purpose of this example is to present how to model and assess creativity within a game called *Physics Playground* (Shute & Ventura, 2013).

UNCORRECTED PROOFS

Physics Playground is a two-dimensional physics game designed to assess and support players' conceptual physics understanding, persistence, and creativity. The goal of the game is to draw objects on a screen using a mouse and colored markers to help the green ball reach the red balloon. Created objects then come to life in line with the laws of physics. Different levels in the game require players' application of their emergent conceptual understanding of Newton's three laws of motion, as well as mass, gravity, conservation of energy and momentum (Shute et al., 2013). Players are provided with a simple tutorial at the beginning of the game to learn how to create what are called "agents of force and motion" (i.e., simple machines) to help them solve different puzzles. The agents of force and motion include ramp, lever, pendulum, and springboard (see Shute & Ventura, 2013, for more details about the game).

Based on the literature, creativity was defined as encompassing three main facets: fluency, flexibility, and originality (Guilford, 1956). *Fluency* refers to the ability to produce a large number of ideas (also known as divergent thinking and brainstorming); *flexibility* is the ability to synthesize ideas from different domains or categories (i.e., the opposite of functional fixedness); and *originality* means that ideas are novel and relevant. There are other dispositional constructs that are an important aspect of creativity, but, due to the nature of the game, we decided to focus on the cognitive skills of creativity. Figure 22.6 shows the CM variables of creativity and a few examples of associated in-game indicators identified as evidence for the CM variables.

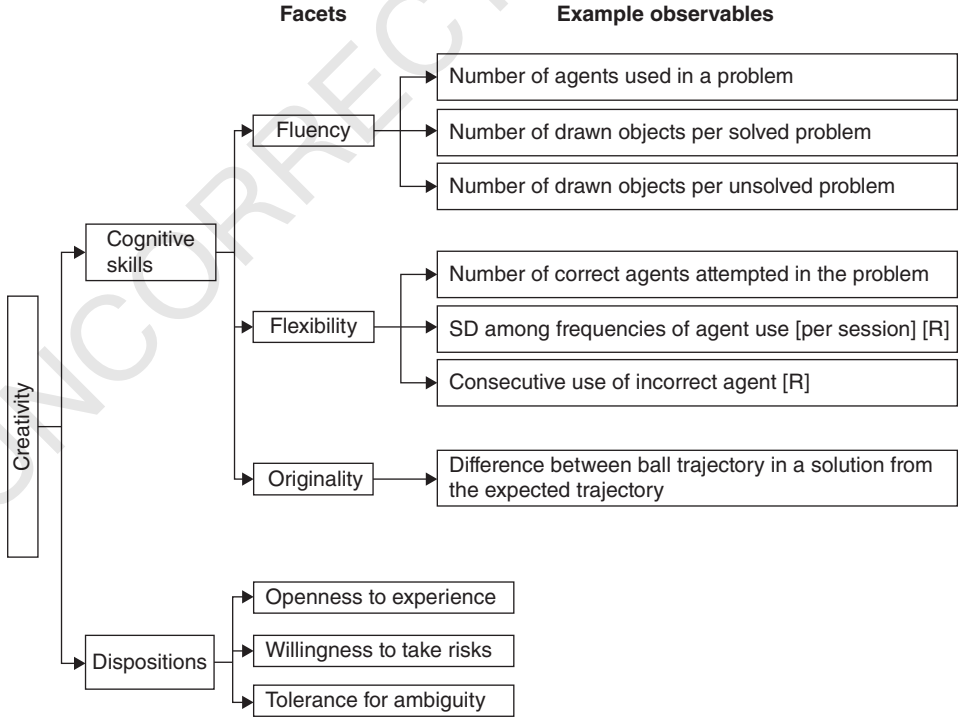

Figure 22.6 CM of creativity and sample indicators; [R] indicates reverse coded.

UNCORRECTED PROOFS

To illustrate how the creativity stealth assessment works in Physics Playground, consider one of the game's 74 levels called "Swamp People" shown in Figure 22.7a, which is a medium-difficult level in the game. In a larger study (Shute et al., 2013) we tested the game with 167 middle school students and found the most common solution to "Swamp People" was to create a ramp from ball to balloon as shown in Figure 22.7b. Less frequently, students created a springboard to solve the level as shown in Figure 22.7c. In just one case, a student used a lever, situated above the alligator's head to solve the problem as shown in Figure 22.7d.

When designing this level, the game designers did not expect anyone to use a lever solution, so the solution shown in Figure 22.7d provides positive evidence for flexibility and originality, with a higher weight for originality compared with solutions shown in Figures 22.7b and 22.7c. Evidence was identified and scored from gameplay and accumulated in the corresponding Bayes net. Each level in the game had its own Bayes net as the levels differed in terms of difficulty as well as set of applicable agents.

Figure 22.8 shows the Bayes net for the "Swamp People" level similar to the Bayes net in Figure 22.5 from the previous example. The upper four nodes represent the overall creativity estimate along with the three main skills (fluency, flexibility, and originality), which are the theoretical constructs on which we are making inferences. The lower nodes represent the gameplay indicators, which are the observables that will be used as the basis for rendering inferences; we only include the indicators for flexibility and originality for illustration purposes.

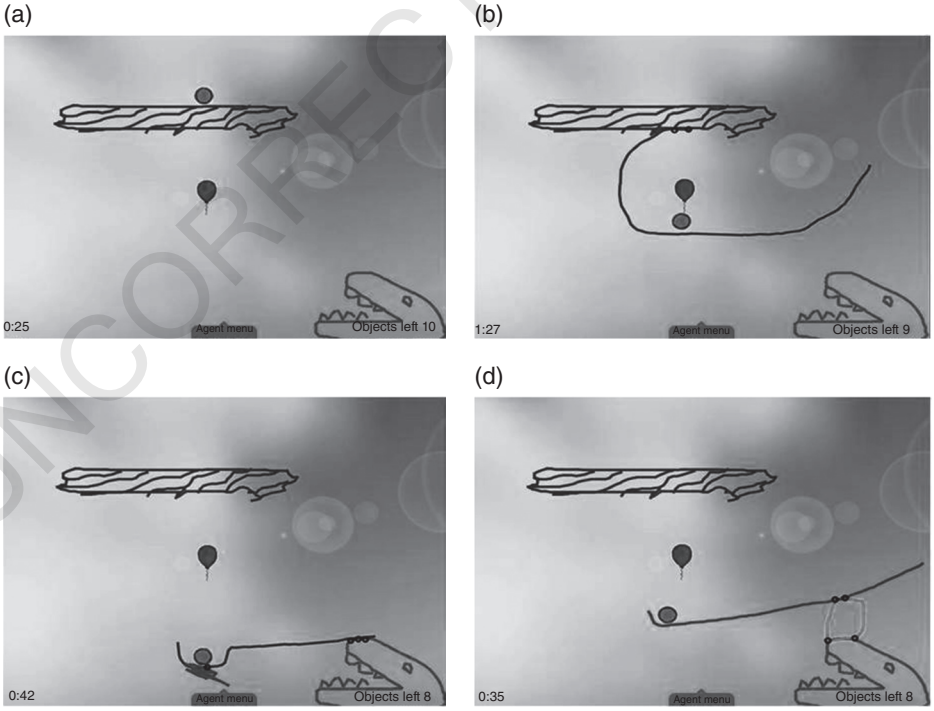

Figure 22.7 "Swamp People" level in Physics Playground (a) with typical ramp solution (b), less common springboard solution (c), and rare lever solution (d).

UNCORRECTED PROOFS

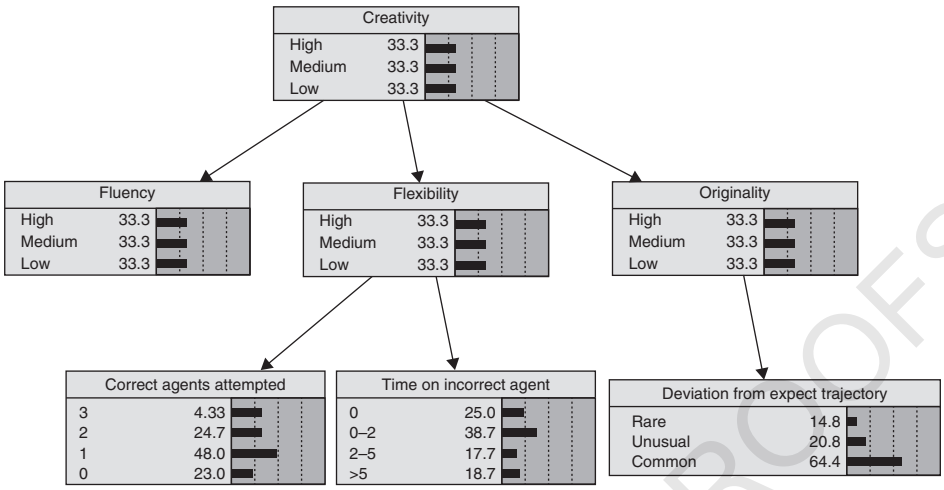

Figure 22.8 Bayes net fragment for creativity for "Swamp People" level.

Before using our data, we initialized this Bayes net with prior probabilities of each indicator node based on an approximately normal distribution except for the probabilities relating to "deviation from expected trajectory" because we assumed that student trajectories would more likely be "common" than "rare" and "unusual." Although prior probabilities came from expert opinions, the estimates become more accurate as more student performance data are entered into the nets. Using the gameplay examples shown in Figure 22.7b, Figure 22.7c, and Figure 22.7d, the estimates for students' creativity are updated. That is, at the end of a level, which occurs if the student successfully solves the problem or leaves the level, data from the log file are analyzed and observables (i.e., indicators) are automatically created, scored, and inserted into the Bayes net.

Consider Student 1 who solved the level with a ramp (Figure 22.7b) as an example. The log file showed that she created a ramp (a common and expected solution) to solve the level after just one failed attempt with a ramp. While this behavior would have a positive impact on the estimate of her emerging physics understanding (specifically related to potential and kinetic energy), it has a low impact on creativity. After her evidence was inserted into the Bayes net, the updated probability distribution in the parent node for creativity suggests that she is likely to be "low" to "medium" in creativity. As shown in Figure 22.9, Pr (Creativity = high | evidence) = 0.20, Pr (Creativity = medium | evidence) = 0.37, Pr (Creativity = low | evidence) = 0.43. More evidence, of course, is needed to increase the confidence of this claim.

In contrast, consider Student 10 who had a different solution **strategy**. According to his log file, he first attempted a pendulum solution for about 47 seconds, which is not an applicable solution for this level (see evidence for "time on incorrect agent" node). Next, he switched to creating a springboard (see Figure 7c). After several failed attempts, he finally succeeded in getting the green ball up to the balloon. The updated Bayes net for this student is shown in Figure 22.10, which displays his estimates for creativity: Pr (Creativity = high | evidence) = 0.63, Pr (Creativity | medium) = 0.32. Pr (Creativity = low | evidence) = 0.05.

UNCORRECTED PROOFS

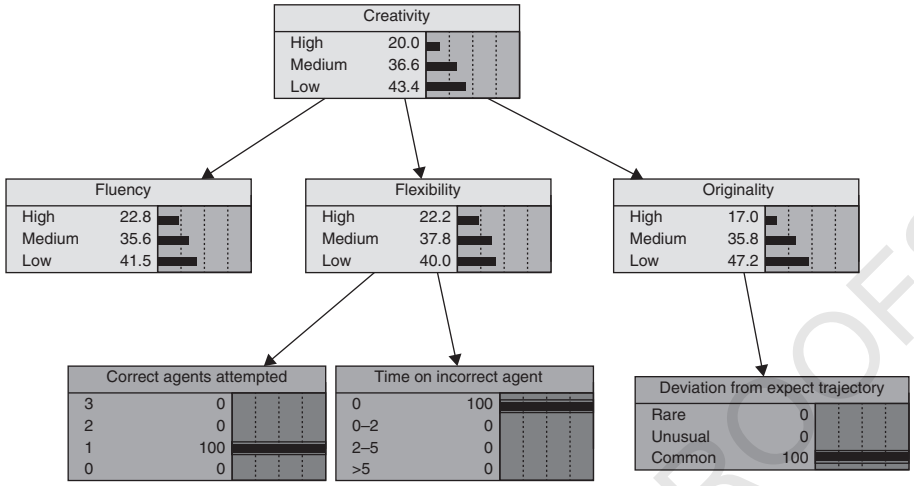

Figure 22.9 Ramp solution (typical) to the "Swamp People" level.

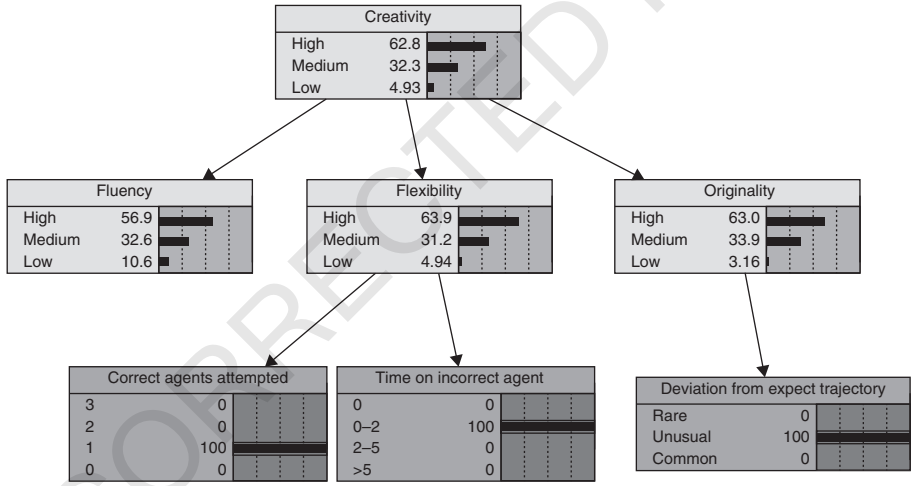

Figure 22.10 Springboard solution (less common) to the "Swamp People" level.

Next, consider Student 3 who tried using a springboard for a while (a viable agent to solve this problem) but then switched to using a lever to solve the problem (see Figure 22.7d), another viable agent. His Bayes net estimates are shown in Figure 22.11: Pr (Creativity = high | evidence) = 0.75, Pr (Creativity = medium) = 0.23, Pr (Creativity = low | evidence) = 0.02. Again, his creation of a lever to solve the problem was unique, impacting the originality facet quite strongly. More data are needed to see if the claim of being highly creative holds across multiple levels in the game.

Finally, consider Student 4 who continued to use a pendulum (i.e., an inapplicable agent for this level) to attempt to solve the level. She failed to solve it and left the level after playing for 6 minutes and 53 seconds; note that spending such a long time on an unsolved level positively impacts her persistence estimate. Because she appeared

UNCORRECTED PROOFS

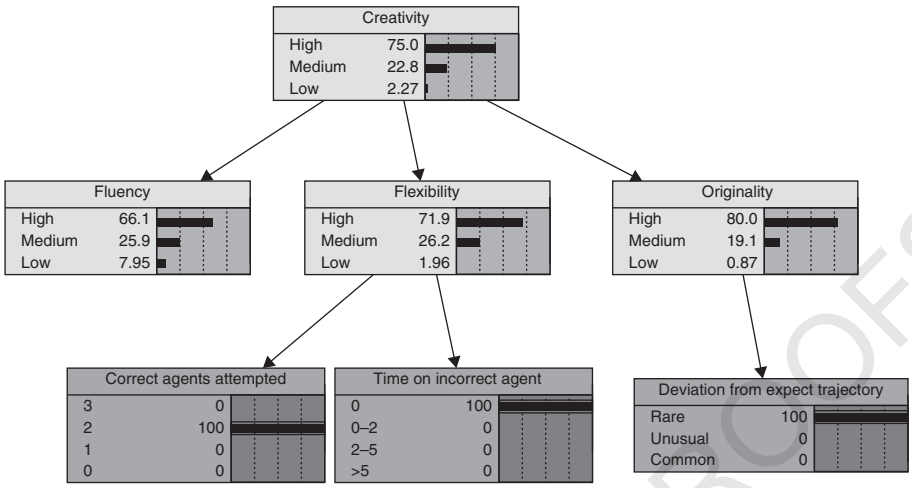

Figure 22.11 Lever solution (rare but effective) to the "Swamp People" level.

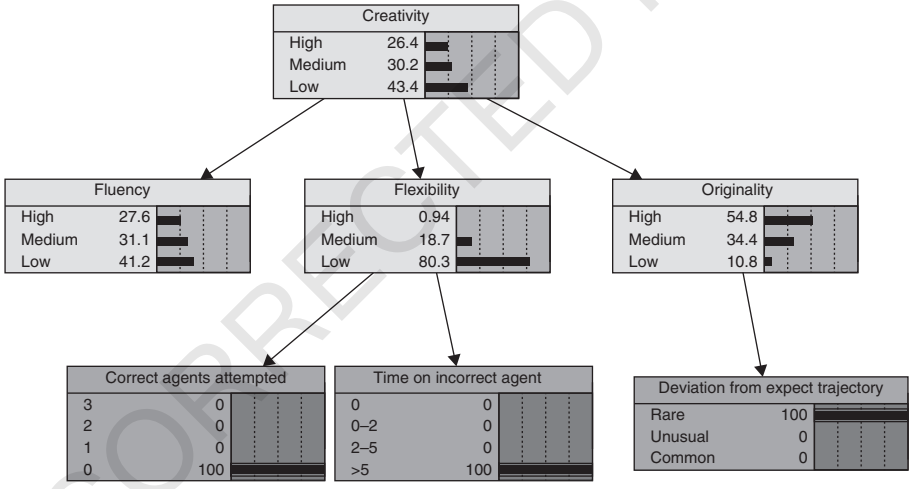

Figure 22.12 Pendulum solution (unusual and ineffective) to the "Swamp People" level.

fixated on creating only pendulums, this led to a low flexibility estimate. Although the ball trajectory in her solution attempts deviated quite a bit from the expected trajectory (positively impacting the originality facet), the ball never hit the red balloon so her solution attempts failed to solve the problem, which is a critical criterion in judging creativity. The actions she took on this level reduced her creativity estimates in the Bayes net, which is shown in Figure 22.12: Pr (Creativity = high | evidence) = 0.26, Pr (Creativity = medium | evidence) = 0.31, Pr (Creativity = low | evidence) = 0.43.

These four examples illustrated how different behaviors in playing the game can be used to infer students' level of creativity for just one level in the game and for the overall construct and its skills or facets. We designed our log files so that we could capture players' states relative to each indicator and feed that information to the Bayes nets.

UNCORRECTED PROOFS

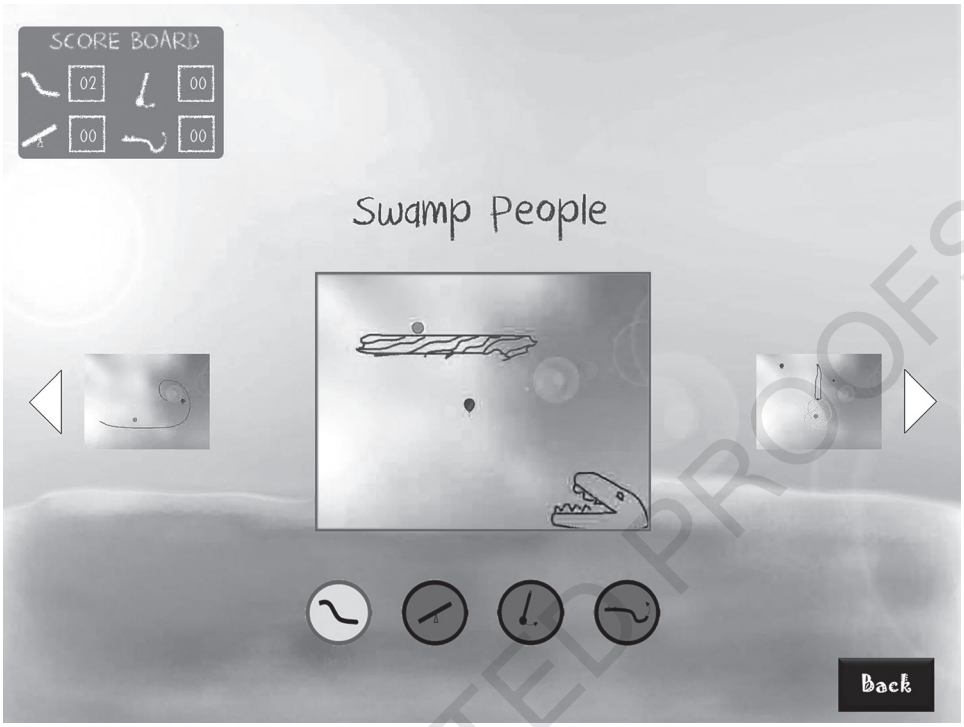

Figure 22.13 Physics Playground dashboard showing score board per agent of force and motion.

Scores are updated immediately once new information enters the network, and they are accumulated over time and gameplay.

In the main interface of Physics Playground (see Figure 22.13), students, teachers, and parents may see the general progress of the player reflected in a score board in the upper-left part of the screen. In the game, successfully solving a level – after repeated attempts – earns the player a silver trophy for the relevant agent (single point). Solving the level "elegantly" with less than three objects earns the player a gold trophy for the level, which is worth double points. Currently, Physics Playground only displays progress on physics understanding, relative to the agents of force and motion, but the same "scoreboard" idea can also be used to present creativity and persistence estimates.

Validation of In-Game Measures

The preceding two examples showed how we adopted best practices of ECD to design stealth assessments embedded in a commercial game (PvZ2) and in a "homemade" game (Physics Playground) to assess two important yet hard-to-measure constructs within two different game environments. Our examples had reasonable CMs and EMs derived from extensive literature reviews, consultation with experts, and observing players engage in gameplay for both problem-solving skill and creativity.

UNCORRECTED PROOFS

The continuous nature of game-based assessments provides for a broader and deeper sampling of relevant evidence than can be obtained from more traditional assessment formats. This, of course, should have positive implications for both reliability and validity. However, it is important to actually subject new stealth assessments to rigorous psychometric scrutiny. That is, the immediate next step needs to address the question of whether or not in-game measures are reliable and, importantly, valid measures of what they purport to measure. We also need to test the degree to which these assessments results can generalize beyond their specific game environments, particularly to real-world settings involving problem-solving skills and creativity. In short, we argue for adopting and adapting best practices for assessment design and validation from standard assessment contexts to game-based assessment contexts (for more on lessons learned and best practices see Wang, Shute, & Moore, in press).

As mentioned earlier, in PvZ2 we conducted a pilot study, which was conducted to determine playability and also to examine preliminary validity evidence. Specifically, we collected gameplay data from 10 undergraduate students who played PvZ2 for about two hours who also completed an external measure of problem-solving skill called MicroDYN, which had been validated previously (Wustenberg et al., 2012). To evaluate correlational patterns between stealth assessment estimates of problem solving and the external measures we first reduced the probability estimates of the overall problem-solving node (e.g., "high," "medium," and "low") to a single number. To do this we assigned numeric values to the three states (+1, 0 and −1) and computed the *expected a posteriori* (EAP) estimate accordingly as $1^*P(\text{High}) + 0^*P(\text{Med}) - 1^*P(\text{Low}) = P(\text{Cij} = \text{High}) — P(\text{Cij} = \text{Low})$ where Cij is the value for Student i on Competency j. This results in a competency estimate scale ranging from −1 to 1. As mentioned earlier, the correlations between our problem-solving EAP estimate from gameplay and the external measures of problem solving (Raven's progressive matrices and MicroDYN scores) were significant. We next want to conduct a larger validation study with middle school students to examine reliability, validity, as well as near and far transfer.

With Physics Playground, our initial evaluation studies focused on establishing the reliability and validity of the physics and persistence stealth assessment measures, both of which were validated against appropriate external measures (Shute et al., 2013; Ventura & Shute, 2013). While we have not fully validated the creativity stealth assessment, we plan to do so against some well-established creativity tests such as *Torrance Tests of Creative Thinking* (Torrance, 1974), *Wallach and Kogan's Creativity Tests* (Wallach & Kogan, 1965), and *Guilford's Alternative Uses Task* (Guilford, 1967).

This brings up another issue that educators and researchers interested in these types of constructs need to address. Because both examples we presented were measured in a particular context (i.e., an interactive game environment), the conventional way of validating new assessments (i.e., testing correlations with existing measures) may not be the most reasonable method for validating hard-to-measure assessments in games.

For example, when validating our in-game persistence measure in Physics Playground, the external measures commonly used to assess persistence are self-reports on questions with 5-point Likert-scales. Given the aforementioned problems with this format, we expected – and indeed obtained – a small, not significant correlation between in-game performance and relevant items from the *International Personality Item Pool* ($r = .01$). When we correlated our in-game persistence measure with a more appropriately

UNCORRECTED PROOFS

matched performance-based measure, however, the correlation was, perhaps not surprisingly, much higher ($r = .51$; $p < .01$); for more details see Ventura and Shute (2013).

The external measure MicroDYN that we used in the PvZ2 pilot study is another example of a performance-based measure. Each item in this problem-solving assessment represents a real-world system, requiring participants to figure out causal relations among different variables, and then manipulate the variables to control the system in specific ways. In our planned larger validation study, we expect that the observed correlations between our PvZ2 stealth assessment measures of problem solving and MicoDYN will extend to other external problem solving measures.

In addition to the possibility for improved reliability and validity that game-based assessments may offer, another positive feature of game-based assessment is that one source of evidence can inform multiple competencies. For instance, consider a case where a person spends a longer-than-normal amount of time on one particular level in Physics Playground. This evidence, coupled with similar cases of spending a long time on tough levels, would positively influence the persistence score, negatively influence the flexibility score, and suggest some issues that may exist regarding the student's understanding of associated physics principles. In a related vein, Almond, Kim, Velasquez, and Shute (2014) noted that the hard-to-measure construct of creativity proved to be the most difficult of the three skills (physics understanding, persistence, and creativity) that could be measured in Physics Playground. Open-ended game levels, like we presented earlier, permit multiple solution paths and are thus good vehicles for measuring creativity.

One final thing to bear in mind when considering the use of games for assessment is how the game goals are framed. For instance, in one study with Physics Playground (Kim, 2014), during data collection, students were told that the person who completed the most levels would get an extra gift card. This one statement may have steered many students towards efficient rather than creative solutions. Other instructions would likely foster other gaming goals and behaviors.

Discussion

Researchers and practitioners are beginning to embrace the idea of using games as a medium to measure and enhance learning and the literature in this new area is moving from rhetoric to more rigorous and systematic analysis of what works, for whom, when, how, and why. Games are obviously not a panacea for educational woes but they are likely superior to traditional methods (e.g., multiple choice assessments, self-reports) when measuring twenty-first-century competencies. In this chapter, we discussed why some of the important twenty-first-century skills are hard to measure and presented our approach to measuring these skills. We believe that performance-based stealth assessments embedded in games provide a meaningful context in which to measure many of these important skills such as creativity and problem solving.

We specifically showcased here two examples of how to model and assess some hard-to-measure but important constructs using a commercial game (i.e., Plants vs. Zombies 2 to measure problem-solving skills) and a "homemade" game (i.e., Physics

UNCORRECTED PROOFS

Playground to measure creativity). The engaging game environments elicit players' target competencies but without any of the test anxiety baggage. Both games are highly engaging, which can lead to greater validity of the assessment. Valid estimates of players' competencies provide a solid basis for delivering on-target and ongoing feedback to players that can foster learning, which is difficult to achieve in traditional classrooms through traditional tests.

In the Physics Playground project, we created many open-ended levels in the game based on the task models, which made it possible for students to demonstrate different aspects and levels of creativity. We customized the log files in Physics Playground so that they captured all observable performance data necessary to input to the Bayes nets, which made the data analysis process efficient.

In the PvZ2 project, we are working with a ready-made commercial game and, instead of creating tasks, we identified indicators of problem solving from each game level. In collaboration with the GlassLab technical team and with permission from Electronic Arts, we were able to modify the game code (minor changes) and log files (major changes) to ensure the capture of necessary data (e.g., x/y coordinates of where different plants were placed, timestamps of all actions) to input to the Bayes nets.

Based on our experiences to date in designing valid assessments of hard-to-measure constructs in game environments, we feel that it is most efficient and effective to bring together educators, game designers, and assessment experts to work together from the onset. This type of heterogeneous team is a critical part of creating an effective learning ecosystem. Having a shared understanding of educational and gaming goals is key to moving forward with the design of engaging, educational games.

Acknowledgments

We would like to thank the Bill and Melinda Gates Foundation (#OPhysics Playground1035331) who supported the development and evaluation of Physics Playground, as well as the Physics Playground team at Florida State University – Matt Small, Matthew Ventura, Russell Almond, Yoon Jeon Kim, and Weinan Zhao – for working their magic into this project. We are also very grateful to the staff at GlassLab who supported our work assessing problem solving in Plants vs. Zombies 2 – specifically Jessica Lindl, Liz Kline, Michelle Riconscente, Ben Dapkiewicz, and Michael John. Finally, we'd like to thank the two reviewers of this chapter – André A. Rupp and Janice Gobert – for their sage feedback on parts of this chapter.

References

Abe, J. A. (2005). The predictive validity of the five-factor model of personality with preschool age children: A nine-year follow-up study. *Journal of Research in Personality*, *39*, 423–442.

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

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). Rejoinder to comments on task features in simulations and games. *Measurement: Interdisciplinary Research and Perspectives*, *12*(3), 118–124. doi: 10.1080/1536637.2014.939628

UNCORRECTED PROOFS

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223–237.

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

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, *1*(5). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1671/1509

Baker, R. S. (2005). *Designing intelligent tutors that adapt to when students game the system*. Doctoral dissertation. CMU Technical Report CMU-HCII-05-104.

Bransford, J. D., Brown, A. L., Cocking, R. R., Donovan, M. S., & Pellegrino, J. W. (Eds.) (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

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

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

Dede, C. (2005). Planning for neomillennial learning styles. *EDUCAUSE Quarterly*, *28*(1), 7–12.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

Eisenberger, R. (1992). Learned industriousness. *Psychological Review*, *99*(2), 248–267.

Eisenberger, R., & Leonard, J. M. (1980). Effects of conceptual task difficulty on generalized persistence. *American Journal of Psychology*, *95*(2), 285–298.

Friedman, T. (2005). *The world is flat: A brief history of the globalized world in the twenty-first century*. London, UK: Allen Lane.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)-Theoretical and Practical Computer Applications in Entertainment*, *1*(1), 20–24.

Gee, J. P. (2007). Games and learning: Issues, perils and potentials. In J. P. Gee (Ed.), *Good video games and good learning: Collected essays on video games, learning and literacy* (pp. 129–174). New York, NY: Palgrave/Macmillan.

Gobert, J., Sao Pedro, M., Raziuddin, J., & Baker, R. (2013). From log files to assessment metrics for science inquiry using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, *53*, 267–293.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Hunt, E., & Minstrell, J. (1996). Effective instruction in science and mathematics: Psychological principles and social constraints. *Issues in Education*, *2*(2), 123–162.

Ito, M., Baumer, S., Bittanti, M., Boyd, D., Cody, R., Herr-Stephenson, B., … Tripp, L. (2010). *Hanging out, messing around and geeking out: Kids living and learning with new media*. In J. D. & C. T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying fusion model to LanguEdge assessment. *Language Testing*, *26*(1), 31–73. doi: 10.1177/0265532208097336

Jirout, J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental Review*, *32*(2), 125–160.

Kim, Y. J. (2014). *Search for the optimal balance among learning, psychometric qualities, and enjoyment in game-based assessment* (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

UNCORRECTED PROOFS

Kim, Y. J., & Shute, V. J. (2015). Opportunities and challenges in assessing and supporting creativity in video games. In G. Green & J. Kaufman (Eds.), *Video games and creativity* (pp. 100–121). San Diego, CA: Elsevier.

Lanyon, R. I., & Goodstein, L. D. (1997). *Personality assessment* (3rd ed.). New York, NY: Wiley.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, MA: Cambridge University Press.

Lubart, T. I. (1994). Creativity. In R. J. Sternberg (Ed.), *Thinking and problem solving* (pp. 289–332). San Diego, CA: Academic Press.

McCalla, G. I., & Greer, J. E. (1994). Granularity-based reasoning and belief revision in student models. In J. E. Greer & G. I. McCalla (Eds.), *Student modelling: The key to individualized knowledge-based instruction* (pp. 39–62). NATO ASI series F, Computer and systems sciences, Vol. 125, New York, NY: Springer-Verlag.

Minstrell, J. (2001). The role of the teacher in making sense of classroom experiences and effecting better learning. In D. Klahr and S. Carver (Eds.), *Cognition and instruction: 25 years of progress* (pp. 121–150). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1994). Evidence and inference in educational assessment, *Psychometrika*, *59*(4), 439–483.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

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

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, *93*, 116–130.

O'Connor, M., & Paunonen, S. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*, 971–990.

Partnership for 21st Century Learning (2012). http://www.p21.org

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes: Measures of social psychological attitudes* (Vol. 1, pp. 17–59). San Diego, CA: Academic Press.

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

Poropat, A. E. (2009). A meta-analysis of the Five-Factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322–338.

Prensky, M. (2005). Computer games and learning: Digital game-based learning. In J. Raessens, & J. Goldstein (Ed.), *Handbook of computer game studies* (pp. 97–122). Cambridge, MA: The MIT Press.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*, *49*(3), 363–393.

Raven, J. (1962). *Advanced progressive matrices: Sets I and II*. London, UK: H. K. Lewis.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1–48.

Razzouk, R., & Shute, V. J. (2012). What is design thinking and why is it important? *Review of Educational Research*, *82*(3), 330–348.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, *8*(4), 1–47.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guildford Press.

UNCORRECTED PROOFS

Salen, K., & Zimmerman, E. (2005). Game design and meaningful play. In J. Raessens & J. Goldstein (Eds.), *Handbook of computer game studies* (pp. 59–80). Cambridge, MA: MIT Press.

Schmitt, N. (1994). Method bias: The importance of theory and measurement. *Journal of Organizational Behavior*, *15*, 393–398.

Shiner, R. L., Masten, A. S., & Roberts, J. M. (2003). Childhood personality foreshadows adult personality and life outcomes two decades later. *Journal of Personality*, *71*, 1145–1170.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning and Media*, *1*(2), 1–11.

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., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it – or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education*, *18*(4), 289–316.

Shute, V. J., Ke, F., & Wang, L. (in press). Assessment and adaptation in games. To appear in P. Wouters & H. van Oostendorp (Eds.), *Techniques to facilitate learning and motivation of serious games*. New York, NY: Springer.

Shute, V. J., Kim, Y. J., & Razzouk, R. (2010). *ECD for dummies*. Retrieved from http://myweb.fsu.edu/vshute/ECD%20for%20Dummies/ECD%20for%20Dummies.swf

Shute, V. J., Lajoie, S. P., & Gluck, K. A. (2000). Individualized and group approaches to training. In S. Tobias, & J. D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military* (pp. 171–207). New York, NY: Macmillan.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning*, *8*(2), 137–161.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570–600). New York, NY: Macmillan.

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

Shute, V. J., & Torres, R. (2012). Where streams converge: Using evidence-centered design to assess Quest to Learn. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 91–124). Charlotte, NC: Information Age Publishing.

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

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge/Taylor & Francis.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, *80*, 58–67. doi: 10.1016/j.compedu.2014.08.013

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research*, *106*, 423–430.

Shute, V. J., & Wang, L. (2015). Measuring problem solving skills in Portal 2. In P. Isaias, J. M. Spector, D. Ifenthaler, & D. G. Sampson (Eds.), *E-learning systems, environments and approaches: Theory and implementation* (pp. 11–24). New York, NY: Springer. doi 10.1007/978-3-319-05825-2_2

UNCORRECTED PROOFS

Steinberg, L. S., & Gitomer, D. H. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science*, *24*(3), 223–258.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal*, *18*(1), 87–98.

Torrance, E. P. (1974). *Torrance tests of creative thinking: Norms-technical manual*. Lexington, MA: Ginn and Company.

Trapmann, S., Hell, B., Hirn, J. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Journal of Psychology*, *215*, 132–151.

Trilling, B., & Fadel, C. (2009). *Twenty-first-century skills: Learning for life in our times*. San Francisco, CA: Jossey-Bass.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

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

Vygotsky, L. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.

Walberg, H. J., & Stariha, W. E. (1992). Productive human capital: Learning, creativity, and eminence. *Creativity Research Journal*, *5*, 323–340.

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

Wang, L., Kim, Y. J., & Shute, V. J. (2013). "Gaming the system" in Newton's Playground. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 85–88). Berlin, Germany: Springer-Verlag.

Wang, L., Shute, V. J., & Moore, G. (in press). Lessons learned and best practices of stealth assessment. To appear in the *International Journal of Gaming and Computer Mediated Simulations*, Guest Editor: R. N. Landers.

Wustenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving – more than reasoning? *Intelligence*, *40*, 1–14.

Zimmerman, B. J., & Schunk, D. H. (2001). *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed.). Mahwah, NJ: Erlbaum.

UNCORRECTED PROOFS