Games for Assessment

Valerie J. Shute & Chen Sun

Abstract

An accurate and dynamic game-based assessment (GBA) can be used to monitor ongoing learning of a range of important competencies, and can also enable timely scaffolding for learners—i.e., specific learning supports provided at the right time—thus providing adaptivity (e.g., Plass et al., 2015). One of the best affordances offered by such games is that they are highly engaging—which stands in contrast with most traditional assessments. Ultimately, the goal of GBA research is to inform the design of high-quality, dynamic games that are engaging, adaptive to individual needs, and can support learning (Shute, Ke, & Wang, 2017). In this chapter, we begin with a comprehensive review of GBA, focusing on recent studies related to its design, validity, implementation, along with its relationship with learning. Additionally, we elaborate on how to design, embed, implement, and validate a GBA in a popular game (*Plants vs. Zombies 2*) to assess problem solving skills. We end with a discussion about theoretical and practical implications of GBA, as well as suggestions for future research.

Introduction

The U.S. Department of Education recently blueprinted educational technology (USDOE, 2016). Technology should not only support teaching and learning, it should also help to innovate assessment relative to measuring both cognitive (i.e., knowledge and skills) and non-cognitive (e.g., affective) outcomes. Our premise in this chapter is that well-designed games—educational and commercial—represent a promising vehicle for not only promoting students’ interest and
engagement in various fields, but also supporting active learning and assessment of a range of important competencies.

Over the past couple of decades, a wide array of games have emerged, supporting the development of various competencies, including visual-spatial abilities and attention (Green & Bavelier, 2007, 2012; Shute, Ventura, & Ke, 2015), cognitive shifting skills (Parong et al., 2017), persistence (Ventura, Shute, & Zhao, 2013), creativity (Kim & Shute, 2015a; Jackson et al., 2012), civic engagement (Ferguson & Garza, 2011), and academic content and skills (Coller & Scott, 2009; DeRouin-Jessen, 2008; Dugdale, 1982; Habgood & Ainsworth, 2011; for reviews, see Clark, Tanner-Smith, & Killingsworth, 2016; Tobias & Fletcher, 2011; Wilson et al., 2009; Young et al., 2012). Moreover, game playing is popular across all gender, ethnic, and socioeconomic lines (Entertainment Software Association, 2016). Core game features (e.g., authentic problem solving, adaptive challenges, and ongoing feedback) that are developed in line with various learning theories can engage students affectively, behaviorally, cognitively, and socio-culturally (Plass, Homer, & Kinzer, 2015). For example, leveraging constructivism (Piaget, 1973) and situated learning (Lave & Wenger, 1991) can create environments that foster the positive experience of flow (Csikszentmihalyi, 1990) and cultivate mindsets that promote effort-driven, challenge-centered competency development (e.g., Yeager & Dweck, 2012).

Therefore, we focus on GBA as a type of assessment where a well-designed digital game serves as the vehicle to measure the degree to which learners are acquiring targeted knowledge and/or skills, and support learning processes and outcomes to fulfill educational objectives. But what is GBA? Some researchers operationalize GBA as external assessments (before and after gameplay) to find evidence of learning as a function of playing a game (All, Castellar, & Van Looy, 2016; Clark et al., 2016). Others operationalize it as information captured directly in the
game to inform learning (e.g., de Klerk, Veldkamp, & Eggen, 2015; Shute, Wang, Greiff, Zhao, & Moore, 2016). Mislevy et al. (2014) categorized three forms of GBA: (a) student products that are external to the game (e.g., presentations and reports), with raters judging the quality of the products; (b) assessment items that are preprogrammed into games, which may range from simple math problems to complex tasks; and (c) data streams generated throughout gameplay that are used as the basis to identify and score as evidence for assessment (e.g., stealth assessment). Stealth assessment refers to evidence-based assessments that are woven directly into the game environment (Shute, 2011). During gameplay, students produce rich sequences of actions while performing complex tasks, drawing on the very competencies that we want to assess. Evidence needed to assess the skills is thus provided by the players’ interactions with the game itself (i.e., the processes of play, captured in the log files). Stealth assessment uses Evidence-Centered Design (Mislevy, Steinberg, & Almond, 2003) to create relevant conceptual and computational models that are seamlessly embedded into the game so that knowledge and/or skills can be assessed without being noticed by students (Shute & Ventura, 2013). The term and technologies of stealth assessment are not intended to convey any type of deception but rather to reflect the invisible capture of gameplay data, and the subsequent formative use of the information to help learners (and ideally, help learners to help themselves).

In both formal (e.g., school classroom) and informal settings (e.g., after school programs), the games that we’re focusing on in this chapter are interactive, digital games that support learning and/or skill acquisition (Shute, 2011). According to Facer (2003), good games are engaging. They promote full absorption within an activity using age-appropriate challenges and intrinsically motivating objectives. Assessment within such games requires not only data collection and analysis, but may also include meaningful data interpretation, along with
consequential actions taken based on the interpretation to achieve learning objectives (Shute & Ventura, 2013).

Because this type of GBA is based on gameplay performance, students’ interactions with games are recorded as interrelated data points, each of which provides specific evidence for learning (DiCerbo, Shute, & Kim, 2017; Levy, 2014; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Also, GBA provides ongoing assessment based on a continuous stream of data rather than discrete data characterized by standardized tests. As a result, with GBA educators can monitor students’ learning progression over time (Shute, Leighton, Jang, & Chu, 2016). Furthermore, because the assessment is embedded deeply in games, students do not notice they are being assessed (Delacruz, Chung, & Baker, 2010; Shute, 2011). Thus, GBA can be used to assess what cannot be easily measured via short, summative paper-and-pencil tests, and can save time that would normally be used to administer and score tests—so that more time may be devoted to improving learning (Shute, Leighton et al., 2016). Finally, GBA can be used formatively to not just measure but also support learning (Delacruz et al., 2010; Shute, Leighton et al., 2016).

In the following section, we review the literature on game-based assessment, then provide an example of GBA using the game Plants vs. Zombies 2 (PvZ2; Electronic Arts, 2013).

**Literature Review**

GBA, as defined in this chapter, has formative functionality—i.e., it is used for assessing and supporting learning. The rise of such assessments is credited to advances in technologies, the learning sciences, and measurement methodologies (Leighton & Chu, 2016; Shute, Leighton et al., 2016; Timmis, Broadfoot, Sutherland, & Oldfield, 2016). In addition, because games are intended to be engaging, they provide rich and interesting environments for students to
experience and/or explore (Clarke-Midura & Dede, 2010; Gee, 2005)—which contrasts with typical assessments.

How can we accurately assess students’ evolving knowledge, skills, and other attributes via gameplay? Assessing students’ interactions with a game requires the use of a principled assessment design framework. There are several major frameworks from which to choose (see Shute, Leighton et al., 2016). In addition to establishing the infrastructure of the assessment (e.g., competency, evidence, task, and assembly models), designers and researchers need to ensure the psychometric quality of the assessment relative to reliability, validity, and fairness (DiCerbo et al., 2017; Mislevy et al., 2014).

The most commonly-used assessment design framework that is applicable to and suitable for GBA is Evidence-Centered Design (ECD; Mislevy et al., 2003). In a nutshell, ECD frames how to design assessments that can elicit valid evidence to support intended claims. It guides designers to specify targeted competencies, observables that can reveal competencies, and tasks with which students interact. ECD is particularly suitable for GBA design. First, data generated via gameplay are usually multivariate (de Klerk et al., 2015; Levy, 2013). By establishing multivariate competencies in the competency model (i.e., the unobservable or latent variables), researchers can determine the associated behavioral evidence (i.e., the observable variables) and specify the task features to elicit those behaviors, along with values assigned to those behaviors (Mislevy et al., 2003). To extract relevant data from gameplay, it is important to (a) identify relevant types of student-task interactions that provide explicit links between behaviors and competencies; (b) decide on the granularity of the observables to be collected; and (c) choose the appropriate statistical model for accumulating and interpreting evidence (Levy, 2013).
Choosing the right statistical model relates to the second advantage of using ECD; that is, drawing valid inferences of students’ competency states. Bayesian networks (BNs) are the most frequently used statistical model in the ECD framework (de Klerk et al., 2015; Mislevy et al., 2014), where a BN is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. BNs generate conditional probabilities of students’ competencies with graphical representation of the statistical relationship(s) between targeted competency variables and associated indicators. Additionally, BNs are dynamic in updating beliefs about students’ competencies (Mislevy et al., 2014), so that ECD can produce real-time data across time, enabling profiles of learning progression (Shute, 2011; Shute, Leighton et al., 2016).

Some examples of GBA for cognitive skills using an ECD framework include the measurement of scientific inquiry skills (Clarke-Midura & Dede, 2010; Baker, Clarke-Midura, & Ocumpaugh, 2016), systems thinking skills (Shute, Masduki, & Donmez, 2010), creativity (Kim & Shute, 2015a), and problem solving skills (Shute, Wang et al., 2016). Additionally, GBA using ECD is well suited to assess content knowledge in various domains, such as mathematics (Delacruz et al., 2010), urban planning (Rupp, Gushta, Mislevy, & Shaffer, 2010); physics (Shute, Ventura, & Kim, 2013), and biology (Conrad, Clarke-Midura, & Klopfer, 2014; Wang, 2008). It is worth noting that Conrad and colleagues (2014) developed a modified version of ECD (i.e., experiment centered design, or “XCD”) in an online game, specifically for students to conduct scientific experiments to answer questions and input answers, in either open-ended or close-ended contexts.

Leighton and Chu (2016) similarly envisioned a new design framework which integrated ECD with the Cognitive Diagnostic Assessment system (CDA, Embretson, 1998) to offset each
other’s shortcomings. CDA is a framework that focuses on measuring students’ cognitive abilities via items that have been designed to measure, based on theories and models of cognition, specific knowledge structures and processing skills in students to provide information about their cognitive strengths and areas for improvement (Leighton & Chu, 2016). The authors discussed the similarities, differences, and challenges of the two design frameworks. They concluded that the new combined framework—in the hands of learning scientists and subject matter experts—could help identify the most relevant information as evidence, and establish a widely applicable, socioemotional-cognitive assessment model.

In addition to existing assessment design frameworks, there are some homemade frameworks that target a specific game type and/or domain. For instance, Nelson, Erlandson, and Denham’s (2011) framework was designed for the genre of massively multiplayer online virtual games. They identified three primary sources for data extraction: (a) players’ location and movement patterns in the game; (b) interactions with various objects; and (c) the type, content, and purpose of communication activities. They illustrated how to utilize each data source using various virtual games as examples, regarding the kinds of behavioral data to look for and how to interpret the data. Also, they stressed that game-player interactions normally involve at least two of the three sources. Thus, tracking and analyzing data from different sources simultaneously and continuously can enable timely feedback during game play as well as post hoc analyses.

Another framework for 3D GBA has been used in the military. This framework includes ontology creation and Bayesian networks (see Koenig, Lee, Iseli, & Wainess, 2010). Ontology creation involves defining the domain in terms of its elements and relationship(s) between and within elements. BNss are employed to model the relationships. The researchers tested their framework within a fire-fighting game consisting of ten scenarios. After comparing the in-game
estimates with human ratings, per scenario, the researchers reported that the estimates derived from the BNs diverged from expert ratings in several scenarios, but on average, the agreement appeared to be reasonable—at around 58%. They attributed the divergence to either the quality/robustness of the BNs, or to inconsistencies in human ratings over time.

The next section examines general properties of GBA (e.g., validity, learning support, and factors influencing GBA quality).

**Validity of Game-based Assessment**

As with the design and development of any assessment, it is necessary to validate GBA. To accomplish this, some researchers have examined the correlation between in-game measures and external measures, while others have converted existing summative tests to GBA.

**Correlation between in-game and external measures.** In a study using Physics Playground to measure and support physics understanding, Shute et al. (2013) reported significant correlations between the in-game measures related to learning physics (e.g., number of gold and silver trophies obtained, time on task) and external learning outcome scores on a qualitative physics test, suggesting convergent validity. Similarly, Delacruz et al. (2010) examined the validity of a puzzle game to teach and assess math. They showed that the math pretest scores predicted game scores, which in turn predicted math posttest scores (controlling for pretest score). In another study examining the development of math abilities using GBA, Roberts, Chung, and Parks (2016) designed a website containing a series of math games for children. The website employs learning analytics to track indicators like correctness of responses while children are playing the games. Learning analytics are intended to gather, measure, analyze, and report data generated by learners to understand and improve learning and associated contexts (SoLAR, 2011). The in-game analytics significantly correlated with scores from a
standardized math test. Finally, using a game to teach middle school students evolution, researchers reported that certain in-game behaviors (i.e., number of times and duration viewing relevant information, number of avatars used, and number of rounds played) correlated with game scores (Cheng, Lin, & She, 2015). Moreover, game scores and posttest scores were significantly correlated.

In addition to subject matter content, researchers have tested GBA validity in relation to other student skills and attributes. For instance, persistence was measured in Physics Playground based on indicators, such as the average time spent on unsolved problems and number of revisits to work on an unsolved problem (Ventura & Shute, 2013). The in-game measures, in turn, significantly correlated with an external measure of persistence (i.e., performance-based measure; see Ventura et al., 2013) as well as to scores on a physics posttest. Along the same line, Dicerbo (2014) modeled persistence relative to two in-game indicators—total time on game quests and number of quests completed. Confirmatory factor analysis showed a good model fit, where the indicators explained a significant portion of the variance, with a reliability of .87. In another study, Shute, Wang et al. (2016) embedded a stealth assessment of problem solving skills in the game Plants vs. Zombies 2. The in-game measures of problem solving skills significantly correlated with two external measures of problem solving (i.e., MicroDYN; Wüstenberg, Greiff, & Funke, 2012 and Raven’s Progressive Matrices; Raven, 1941).

In conclusion, validating GBA with external measures entails two prerequisites: (a) careful selection of in-game indicators for targeted knowledge and/or skills, and (b) use of a well-established (i.e., valid and reliable) external measure of the same construct.

**Mapping game-based assessment to summative, performance-based assessment.** An alternative approach to establishing the validity of GBA involves mapping GBA to a summative,
performance-based assessment. Two recent studies of this type were conducted, both within a vocational education setting. In one study, the original performance-based test requires assessors to assume different roles (e.g., clients), to interact with students and then judge their qualifications to be an information technology communication manager (Hummel, Brinke, Nadolski, & Baartman, 2016). To ensure the content validity of the GBA, the researchers employed the following four steps: (a) identify relevant performance indicators that can be elicited by gameplay; (b) design game tasks; (c) develop instructions for GBA users; and (d) evaluate whether the game tasks map on to the performance metrics that were the target of the original assessment. After implementing the GBA and interviewing assessment experts, the researchers reported that GBA could fully assess 20 out of 32 performance indicators, and partly assess five more indicators, while the rest of the indicators could be assessed face-to-face. The GBA’s main advantage is to avoid inconsistencies and biases that are frequently present in human ratings. Moreover, it saves time in assessment execution and results documentation.

A second study developed an interactive virtual assessment that had been mapped to real-life performance-based test scenarios (de Klerk, Eggen, and Veldkamp, 2016). The real-life test measured students’ abilities to inspect the working conditions and procedures within a confined space, and then respond properly to emergencies. As with the previous example, they first defined performance indicators in the virtual environment. Then experts rated each indicator in terms of its difficulty and evidentiary weight relative to associated knowledge and skills. Based on experts’ ratings, the authors constructed two scoring models for assigning values to the indicators and then transforming the scores to BNs. Finally, they compared the scores generated by the two scoring models with those on a real-life performance-based test. The results indicated that one model estimated students’ qualifications more accurately than the other.
Such mapping methods are relatively scarce. Further studies are needed to provide evidence for its reliability and validity. One question to consider is whether the GBA can be consistently and accurately mapped to the original performance-based assessment. The other question is whether such a mapped GBA can eventually gain recognition and replace the original assessment to serve a high-stakes, summative purpose. Currently, GBA typically serves a formative function—to support learning processes and outcomes, described next.

**Game-based Assessment to Support Learning**

Well-designed GBA can support some degree of learning without explicit instructional support (see Shute et al., 2013). As mentioned earlier, middle school students played Physics Playground for about three hours (across three days), and also completed pre- and posttests on qualitative physics. The students’ in-game performance was assessed via a range of indicators, such as number of solution attempts, time per level, and level of trophy obtained. The results showed small but significant learning gains of physics understanding, measured by pre- and posttests, and in-game measures were significantly correlated with test scores. Moreover, both male and female students significantly improved their physics knowledge as a function of gameplay. Although the males’ incoming knowledge was slightly higher than the females’, their posttest scores were comparable. The researchers concluded that this GBA is fair to use for both male and female students, and in the future, feedback (e.g., explanations and visualizations) can be integrated in the game to facilitate deep learning.

Game-based assessment also can be leveraged to provide various forms of feedback to support learning. To illustrate, a GBA was designed using ECD to measure knowledge related to geology and space science (see Reese, Tabachnick, & Kosko, 2015). The GBA tallied learners’ progression toward the learning goals every 10 seconds. The data were stored in a database and
thereby served as the basis for timely feedback. The feedback (e.g., on-screen scaffolding messages and player dashboards) was integrated directly into the game to facilitate goal achievement. For instance, current point tallies were displayed at all times, and the scaffolding assumed various forms, such as text, pictures, and animations. Scaffolding was presented when learners repeatedly made mistakes. Based on the data, the researchers calculated the rate of learning toward four topics (i.e., mass, heat, radiation, and density) in two samples for generalizability. The results showed the two samples progressed at similar rates in learning mass, heat, and radiation, but differed relative to learning density. Generally, the learning rates for both samples were significantly greater than zero, providing evidence that GBA can facilitate learning.

Arnab and colleagues (2015) utilized learning analytics in a game to assess and support knowledge of first aid techniques among college students. For the in-game measures, students needed to select their answers (i.e., reactions to different scenarios) from several pictorial choices. Whatever choices the students made, correct or incorrect, the associated consequences of the selected action would show up immediately as feedback. In addition, pre- and posttests were implemented to assess students’ knowledge gains. The results showed that the in-game scores predicted posttest scores, and there were significant learning gains. The researchers recommended future research to use in-game measures to predict students’ performance and offer personalized support based on the in-game estimates.

Other researchers have compared two versions of the same game (e.g., feedback present vs. feedback absent) to test if the GBA with embedded feedback is a better design to improve learning compared with GBA without feedback. In one such comparison study, Huang, Huang, and Wu (2014) designed two versions of a math game where second grade students answered
math questions related to buying various goods. One version provided timely feedback (i.e., hints or explicit feedback) when errors occurred, while the other version did not. The results showed that the students who played the game with diagnostic feedback produced significantly higher posttest scores than those who did not receive feedback. The authors concluded that the diagnostic feedback helped students learn from their errors by providing instructional support according to the types of mistakes they made.

Tsai, Tsai, and Lin (2015) similarly investigated the effects of the GBA with immediate elaborated feedback, compared to the version with only verification feedback on supporting middle school students’ acquisition of energy knowledge. The GBA for energy knowledge assumed the form of a tic-tac-toe game where students needed to answer questions. The GBA placed a tick when the answer was correct, or a cross when the answer was incorrect. In the elaborated feedback condition, immediate explanations of answers to questions were provided on screen for students’ reference, in addition to the verification feedback (i.e., ticks and crosses). The researchers found that only the game with elaborated feedback significantly improved knowledge acquisition from pretest to posttest, supporting other studies’ findings that GBA with timely and explanatory feedback facilitates learning.

Examining the effects of three types of formative assessment on learning, Wang (2008) conducted a study with fifth graders in a two-week biology course. Six classes were randomly assigned to one of the three conditions that used different types of formative assessment to support learning. In addition, the researcher administered a pretest and posttest on biology knowledge, and they used different test items for summative and formative assessments.

The first type of formative assessment was a paper-and-pencil test administered at the end of each class, with correct answers given to students as feedback. The second type of
assessment was a web-based test, where students received immediate feedback concerning the correct answer for each of their incorrect responses. The third type was a GBA (i.e., an online multiple-choice quiz game), where students could press certain buttons to receive a hint (e.g., to see others’ choices, such as “80% of test-takers chose A as the correct answer”). However, use of the hint function was limited to prevent its overuse. Using pretest as a covariate, findings showed that the three types of formative assessment significantly influenced posttest scores. Post-hoc analysis showed that posttest scores in the GBA condition were significantly higher than the other two conditions. The researcher contended that students were motivated by the game-like quiz, and tended to actively refer to resources (e.g., learning materials or asking for clarifications from teachers).

**Game-based Assessment to Model Factors Influencing Learning**

In addition to its ability to support learning, GBA can also be used to identify particular factors and patterns that contribute to successful learning. For example, several researchers have recently examined the behavioral patterns related to science learning (Baker et al., 2016). They analyzed scientific inquiry behaviors among middle school students via a virtual environment that provided various science-related scenarios for students to solve. Focusing on students’ final answers as well as their procedures used to conduct scientific tests, the researchers used confirmatory factor analysis to identify 29 behavioral patterns for successful learning that could be generalized across scenarios. In short, students’ final correct answers were predicted by time spent on and the frequency of visits to an information page. The indicators related to successfully identifying causal relationships included: (a) obtaining necessary items for conducting experiments (e.g., water or blood samples), (b) visiting the virtual science lab frequently, and (c) running relevant tests (e.g., blood or DNA test).
Cognitive and noncognitive variables and their relationships to learning were examined and modeled in a study conducted by Shute et al. (2015). The researchers gathered data from middle school students’ playing Physics Playground, and additionally collected data on students’ persistence, incoming physics knowledge, in-game performance (e.g., time on levels, successful and unsuccessful solutions, trophies received), affective states, as well as physics posttest scores. They used structural equation modeling to construct various models to interpret the relationship between learning outcomes and the other variables. The final model demonstrated that pretest scores significantly related to engagement, in-game performance, and to posttest scores. Also, engagement and frustration were two mediating variables between pretest and in-game performance, suggesting the importance of creating adaptive tasks that exceed students’ current proficiency level by just a little. Furthermore, in-game performance significantly influenced posttest scores. The results of the relationships among the different variables pertaining to learning provide implications for instructional support.

Factors Influencing the Quality of Game-based Assessment

The quality of GBA depends on its underlying framework (e.g., ECD) and its psychometric properties (e.g., reliability and validity), as mentioned at the beginning of this review. A few attempts have been made, to date, exploring the factors that affect GBA quality. For instance, changing task variables can affect the psychometric quality of GBA tasks (Almond, Kim, Velasquez, & Shute, 2014). Tasks possess particular features that govern their presentation as well as the associated work product. These features can affect how learners respond to a task, and the evidentiary weight of the responses. For example, consider a math test on addition and subtraction. The format of the test (e.g., multiple-choice or word problem) influences how much information you would get from students’ responses (e.g., correctness of the choice selected or
the whole problem-solving process). In addition, the format may have some unexpected confounds—such as reading ability serving as a potential confound in the solution of math word problems. There are many other variables to consider in designing and before implementing assessment tasks, such as how to design two different tasks that are of the same difficulty, and how to make sure about 50% of test-takers can complete the task correctly. Task variables help researchers and/or designers determine task variants, difficulty, and discrimination; thus interactions between learners and GBA tasks can yield valid evidence to measure targeted competencies.

Kim and Shute (2015b) examined how game design features (i.e., linearity vs. nonlinearity) affect the psychometric properties of the stealth assessment embedded in Physics Playground. Linearity refers to unlocking game levels, whereas nonlinear games offer learners control of the levels they choose to play. In this study, undergraduates in both linear and nonlinear conditions were instructed to obtain as many points as possible (and were also informed that they can score higher by earning gold trophies for elegant/efficient solutions, which count as double the score of silver trophies). To determine validity, the researchers tested the evidentiary weight of in-game indicators on physics understanding in the two conditions. The evidentiary weight of silver trophies significantly differed between linear and nonlinear conditions. Posttest scores significantly correlated with silver trophies in the linear condition, but with gold trophies in the nonlinear condition. The change in validity might be because linearity did not motivate learners to explore the most efficient solution (i.e., gain gold trophies) to the various physics problems, but instead, just to unlock as many levels as possible by gaining silver trophies. Consequently, only students in the nonlinear condition who aimed for optimal solutions significantly improved their physics learning. To test reliability, the researchers used
confirmatory factor analysis to construct the best-fitting model for both conditions. The calculated reliability coefficients are .96 and .92 for the linear and nonlinear conditions respectively, and the two coefficients are comparable to each other. Thus, reliability of the GBA was not affected by whether the game is linear or nonlinear.

Finally, a recent white paper describes the factors influencing the psychometric qualities of GBA designed with ECD (see Mislevy et al., 2014). The researchers argued that high-quality GBA can serve various purposes (i.e., formative, summative, and even large-scale high stakes) as well as provide valuable information about learning for students, teachers, and designers. They used a game called SimCityEDU, created by Glass Lab and its partners, as a running example to show how to ensure the reliability and validity of a GBA. The area of psychometrics concerns the observable evidence that can be identified and extracted from a given work product—i.e., the log file data in this case—to assess unobservable competencies. The most influential psychometric factors related to GBA involve identifying relevant evidence, and selecting measurement models to trace and process the gaming/learning data. Researchers and designers should additionally consider how to interpret the evidence derived from particular gaming situations, design adaptive games to provide optimal learning experiences, and analyze data related to collaborative activities. They introduced a new framework for this called Evidence-Centered Game Design (ECgD): (a) defining targeted real-world competencies, (b) aligning game-world competencies with the real-world ones, (c) integrating formative feedback systems into the games unobtrusively, and (d) engaging in iterative design processes to create engaging games with embedded assessment to support deep learning.
In the next section, we illustrate the application of a specific type of GBA—stealth assessment in Plants vs Zombies 2 (PvZ2) to measure students’ problem solving skills (Shute, Wang et al., 2016).

**Example of a Game-based Assessment**

*PvZ2* is a widely popular 2D game that requires players to strategically guard their houses against zombie invasion. Players manipulate various plants in the battlefield (i.e., the chessboard-like lawn in front of the house) to either attack zombies directly or slow zombies down. When selecting and placing their plants, players need to collect falling suns to earn energy points. *PvZ2* is an appropriate vehicle to embed a stealth assessment measuring problem solving skills. Again, stealth assessment is defined as an evidence-based assessment woven directly and invisibly into the fabric of the learning or gaming environment (Shute & Ventura, 2013) to measure and support learning. The models undergirding stealth assessment are created using ECD. The combination of ECD and stealth assessment makes it possible to build evidentiary arguments about students’ competency levels via three key models—competency model, evidence model, and task model.

The competency model includes claims about competencies (i.e., unobservable variables) to be assessed. The evidence model specifies behavioral evidence (i.e., observable variables) that can be collected and analyzed/scored to support the claims made in the competency model. The evidence model also quantifies the observables by establishing scoring systems to align evidence with claims statistically. For instance, an observable can be indicated as a ratio to represent various levels of a competency, such as “poor” (0-0.25), “ok” (0.26-0.50), “good” (0.51-0.75), and “very good” (0.76-1). Stealth assessment typically employs Bayesian Networks (BNs) to establish statistical relationships among the indicators and the competency variables. The task
model provides templates for the design of tasks that can elicit targeted evidence. Note that when using an existing game with its existing levels, the task model specification isn’t needed.

To design the stealth assessment in *PvZ2*, the researchers first constructed a competency model of problem solving skills based on an extensive literature review. The overarching competency of problem solving skill involves four facets: (a) analyzing givens and constraints of the problem; (b) planning a solution pathway; (c) using tools and resources effectively and efficiently; and (d) monitoring and evaluating progress. Next, the researchers identified in-game indicators (i.e., the observables) associated with each competency variable (i.e., the unobservables), and then assigned values to indicators to reflect students’ performance quality.

For instance, consider the problem solving facet of “using tools and resources effectively and efficiently.” One of the plants in the game is iceberg lettuce, and its function is defensive—to temporarily freeze zombies. Another plant in the game is the snapdragon. Its function is offensive, attacking zombies by breathing fire and burning them. If a player plants iceberg lettuce within a snapdragon’s fire range, its freezing effects will be canceled by the fire. Thus, one indicator (of many) related to using tools effectively is whether the student planted an iceberg lettuce proximal to a snapdragon (i.e., within a 3 × 3 space; see Figure 1). This indicator was scored by calculating the ratio of iceberg lettuces planted near snapdragons divided by the total number of iceberg lettuce planted. In this case, the higher the ratio is, the lower the associated competency level would be. There are four equally divided ratio intervals, from very good (0-0.25), good (0.26-0.5), ok (0.51-0.75) to poor (0.76-1).
Figure 1. Using iceberg lettuce ineffectively in PvZ2.

After establishing the scoring system across all of the indicators per facet, the researchers constructed BNs to represent the statistical relationships between indicators and relevant competency variables for each game level. Individual BNs were constructed per level because each level varies in terms of its difficulty, discrimination, as well as relevant indicators and competency variables. The prior probabilities of problem solving per student is the same—i.e., equal likelihood of being high (33.3%), medium (33.3%), and low (33.3%) (Figure 2). Then, as data are generated by students during gameplay, these probabilities quickly and repeatedly change. Ongoing data (from the indicators) are input to the BNs, and the BNs process the data and update the competency estimates. The estimates will approach students’ true competency level with the influx of gameplay data because BNs dynamically adjust estimates according to students’ real-time performance.
Figure 2. BN example of prior probabilities (Adapted from Wang, Shute, & Moore, 2015)

Figure 3 shows an updated BN, when the player demonstrates poor iceberg lettuce use (shown in node I37). The updated result means there is about 50% chance that the problem solving skill of this player is low.
In addition to ensuring the internal validity of the stealth assessment, the researchers carefully selected two external measures related to problem solving skills (specifically in terms of rule identification and rule application) to test its external validity. Raven’s Progressive Matrices (Raven, 1941) requires students to infer rules from given matrices to fill in one missing piece of information. MicroDYN (Wüstenberg et al., 2012) requires students to recognize relationships among variables and then apply that rule to achieve desired results. The results from a study conducted with about 50 middle school students playing the game and completing the two external measures across three days showed that the stealth assessment estimates of problem solving skill from the game significantly correlated with the two external measures, suggesting construct validity.
The example illustrates the validity of stealth assessment as GBA. The strength of stealth assessment lies in the following: (a) the competency model is built on a conceptual foundation (i.e., resulting from a comprehensive review of the construct in question), (b) the evidence model establishes specific rubrics for scoring in-game performances as well as statistical relationships between the evidence/indicators and what is being assessed; (c) the assessment is seamlessly and directly embedded into the game resulting in the merging of learning and assessment; (d) learning can be supported by providing timely feedback—at various times and grain sizes; and (e) it is able to concurrently assess multi-dimensional competencies. Next we discuss the theoretical and practical implications and limitations of GBA.

**Theoretical Implications**

This chapter highlights the potential of GBA to measure and support learning simultaneously. Students’ learning can be monitored continuously, without disrupting learning processes (DiCerbo et al., 2017; Shute, Leighton et al., 2016). In addition to content knowledge, GBA is well suited to assess complex skills (e.g., problem solving and creativity) that are normally difficult to assess with traditional measures (Clarke-Midura & Dede, 2010; Timmis et al., 2016). GBA allows for the assessment mechanism to be built directly into the game, comprising an integrated design for games and assessment. Researchers and/or designers can thereby ensure the alignment between learning objectives and assessment tasks, enabling the capture of accurate estimates of students’ knowledge, abilities, and attributes from GBA (Ke & Shute, 2015; Plass et al., 2015).

Another way to obtain accurate estimates of competency states and learning is to employ an appropriate statistical methodology to process GBA data. Currently, BNs are popular because they can accommodate a wide range of models (from simple to complex), generate real-time
estimates accurately, and represent statistical relationships graphically and conveniently (Levy, 2016; Mislevy et al., 2014; Kim, Almond, & Shute, 2016). Researchers can extract copious amounts of GBA data from log files or databases. One downside of log files, though, is readability—especially when log files capture lots of data that are both relevant and irrelevant to the research. One solution to this problem is to modify log files such that they capture only specific evidence (see Shute & Wang, 2016). An alternative approach is to develop a generic log file structure that can be applied to different games to handle data storage and extraction conveniently (see Hao, Smith, Mislevy, von Davier, & Bauer, 2016).

**Practical Implications**

An accurate and dynamic GBA can also enable timely scaffolding for learners—i.e., specific learning supports at the right time—thus providing an adaptivity feature in games (Plass et al., 2015; Virk, Clark, & Sengupta, 2015). For example, based on learners’ current competency estimates from their performances, the game can adjust task difficulty, at 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 et al., 2015; Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Shute, Leighton et al., 2016), especially to support struggling learners (Baker et al., 2016). One thing to keep in mind when presenting various forms of feedback to learners is the cognitive load imposed by various representations and information processing requirements (Adams & Clark, 2014; Lee, Plass, & Homer, 2006; Virk et al., 2015). Also, construct-irrelevant variables (e.g., prior gaming experience) should be controlled to reduce disruption to the gameplay experience (Dicerbo et al., 2017).
Additionally, it is important to consider the accessibility of GBA data. Researchers have argued that learners and teachers should have access to diagnostic data – for students to monitor their learning progress, and to help teachers figure out when and how to intervene, as warranted (Clarke-Midura & Dede, 2010; Shute, 2011; Timmis et al., 2016). Ethical issues should also be taken into account (Pardo & Siemens, 2014; Shute, Leighton et al., 2016; Timmis et al., 2016), such as answers to the following questions: How can the student data be protected? Who owns the data and for how long? How can the data be used to best advantage? Lastly, to integrate teaching, learning, and assessment, researchers advocate close collaboration during GBA design among game designers, researchers, psychometricians, subject matter experts, and other stakeholders (Leighton & Chu, 2016; Mislevy et al., 2014; Plass et al., 2015).

**Limitations and Future Research**

There are several limitations of GBA that will need to be addressed in future studies. The first issue concerns what exactly GBA is. Theoretical papers are needed to clearly define it, as well as describe its various types and distinctive features. For example, the boundary between GBA and simulation-based assessment is not clear. Is GBA a subcategory of simulation-based assessment only with higher levels of interactivity (de Klerk et al., 2016), or are the two overlapping (Levy, 2013)?

A second issue concerns the best statistical tools/analyses to be used to collect and process GBA data. Processing massive and complex gameplay data is difficult, especially when the data involve collaborations (Leighton & Chu, 2016; Hao et al., 2016; Nelson et al., 2011). Thus, figuring out how to effectively combine exploratory techniques (e.g., educational data mining) with more conceptual approaches (e.g., ECD) will benefit GBA research. An additional issue concerns reusability and cost-effectiveness (Moreno-Ger, Burgos, Martínez-Ortiz, Sierra,
Building a well-designed GBA is time-consuming and usually domain-specific. Thus, the applicability of one GBA to other games or disciplines remains an under-researched area (Baker et al., 2016; Wang et al., 2015). The last question involves fairness. It is important that a GBA does not favor any particular population (e.g., males vs. females, and gamers vs. non-gamers) and equally benefits every student (Dicerbo et al., in 2017; Kim & Shute, 2015b; Timmis et al., 2016). However, studies on fairness of GBA are sparse.

One of the main affordances offered by well-designed games is that they are highly engaging. Similarly, well-designed GBA are engaging, in addition to being able to render valid and reliable inferences about students’ competencies during gameplay. The vision is to design high-quality, dynamic GBA that is engaging, adaptive to individual needs, and can support learning (Shute, Ke, & Wang, 2017; Shute, Leighton et al., 2016).

References


