Homer, B., Ober, T., & Plass, J. (2018). Digital Games as Tools for Embedded Assessment. In A. Lipnevich & J. Smith (Eds.), The Cambridge Handbook of Instructional Feedback (Cambridge Handbooks in Psychology, pp. 357-375). Cambridge: Cambridge University Press. doi:10.1017/9781316832134.018

16

Digital Games as Tools for Embedded Assessment

Bruce D. Homer, Teresa M. Ober, and Jan L. Plass

With decreased cost and increased portability, digital technologies have become ubiquitous in nearly all aspects of our lives, including education. During the past decade, one of the promising uses of digital technologies in education has been the use of video games for learning. There are now thousands of educational and learning games, ranging from casual games intended to teach simple concepts (or, more commonly, to reinforce existing knowledge) to more complex and involved games intended to teach deeper knowledge, support the development of complex cognitive skills, and change attitudes or increase awareness. Although a majority of the work in this area has investigated learning outcomes, there is also a growing interest in the use of digital games as tools for assessing learners.

Assessment is a critical component of the education process. Ideally, meaningful assessment provides feedback to students, teachers, parents, and administrators that can be used to improve education outcomes or, in the case of standardized assessments, allow learners to be compared with one another. However, both the development and the implementation of traditional methods of assessment (i.e., paper-based testing) require significant time and resources. In part, this has led to passionate critiques of the current state of standardized testing in our schools, and arguments that the time used to prepare for and administer standardized tests is significantly reducing valuable time that could be devoted to teaching students (e.g., Ravitch, 2016).

One possible solution to this problem that has been offered is the use of digital technologies with assessment embedded into the learning process. Combining the learning and assessment processes not only allows for more instructional time but also enables the possibility of more authentic assessment and more informed instruction. This view is reflected in the most recent supplement to the National Education Technology Plan, in which the US Department of Education (2017) argues that digital forms of assessment can help to "reduce the time, resources, and disruption to learning" caused by traditional modes of assessment. Standard paper-and-pencil forms of assessment are outside the learning process; that is, they are added to the instructional process. In contrast, digital technologies, including video games, can have assessments embedded into the learning context. Using assessment that is embedded into digital learning environments can provide educators with insight into what students are actually thinking while engaging in the learning process and can provide near real-time feedback so that appropriate action can be taken in the moment to support students' learning (US Department of Education, 2017).

As we describe later, these advantages can be particularly true for video games, which have assessment as an essential component. However, in spite of the potential of game-based assessment, there is still only a limited body of research exploring the use of games as assessment tools. In this chapter, we consider the ways in which digital games can be used to authentically evaluate learners’ knowledge and skills. Specifically, this chapter aims to accomplish the following goals:

1. Provide a summary of the research to date on the use of games as tools for assessment
2. Describe models of game-based assessment used to evaluate learning in an authentic manner
3. Present examples from the research literature of successful implementation of game-based learning and assessment
4. Make recommendations for advancing the future of game-based assessment for learning.

Assessment Is Integral to Digital Games

In contrast to the limited work on games as tools for assessment, there is now a substantial body of work on how and why digital games can be effective tools for learning. In his influential book, *What Video Games Have to Teach Us about Learning and Literacy*, Gee (2003) argues that video games embody many of the principles of good learning. These include *agency*, or the fact that players have control over their environment in digital games; *well-ordered problems*, which refers to the fact that players typically solve interconnected problems of increasing complexity in video games (or at least in certain genres of video games); and *customization*, which refers to the fact that games can become easier or more difficult based on the success (or failure) of the player in order to keep the player challenged, but still succeeding. Plass, Homer, and Kinzer (2015) similarly suggest that video games have the potential to embody the best practices of different approaches to learning, and argue that to fully understand the educational potential of video games requires adopting an “overarching, learning sciences perspective” that considers cognitive, motivational, affective, and sociocultural factors.

Arguably, the feature of video games that is most relevant for assessment is their affordance to be adaptive and personalized for individual learners. In order to successfully adapt, an accurate estimate of the learner must be created that informs the adaptation. Commercial entertainment games need only model whether or not the player is successful in the game. If “yes” (i.e., if the player is successful), then the game can “level up” (e.g., by increasing speed, decreasing player resources, providing additional obstacles, etc.), and if “no,” then the game can “level down” (e.g., by decreasing speed, increasing player resources, removing obstacles, etc.) in order to keep the player challenged and engaged. For learning games, wider and more complex assessments are required. In a paper arguing that games are the “future of assessment,” Gee and Shaffer (2010) point out that games must be good “assessment engines” in order to be effective tools for learning. The authors claim that video games can effectively evaluate students’ current knowledge and skills as well their broader “twenty-first century skills,” which is then used to adapt the difficulty or content of the learning game.

From very early on, video games have been used to educate and evaluate users’ knowledge in a wide range of domains. For example, classic and popular educational games such as the Oregon Trail, which dates back to 1971, and Where in the World Is Carmen Sandiego?, released in 1985, utilize the game context to promote learning, requiring users to apply their knowledge, monitor their understanding, and solve in-game problems using domain-specific knowledge (e.g., information from history or geography). Within the context of these games (versions of which are still available today), users discover consequences resulting from their choices as the game tracks the users' decisions and progress throughout the duration of game play. To succeed in the games requires users to apply their knowledge of history and geography – success or failure in the game is a direct assessment of users' knowledge in these areas. If they player fails, they are encouraged to "try again" – in other words, to go back and learn the information that is needed in order to succeed.

In this regard, the concept of evaluating and responding to users' performance while playing a game is not novel. Even early computer games were designed with what could be considered to be some sort of basic formative assessment. *Formative assessment* is a technique where a learner's mastery of a concept or skill is regularly evaluated and instruction is adjusted to accommodate their needs (Black & Wiliam, 1998). The immediacy of feedback provided by these early educational games allowed the learner to know whether or not specific concepts or skills had been acquired and, if not, to learn them in order to succeed at the game.

As the technology has developed, so too has the quantity and detail of the data gathered by games. Most games now collect logs of some sort that not only assess whether or not the player succeeded but also collect details about how the player progressed through the game. This information is typically used to inform revisions to the game. For example, if there a spot in the game where many players are dying in a way that was not intended by the game designers (e.g., falling off a cliff because it is not visually clear that the road curves), then the next version of the game can be modified to remove this issue (e.g., putting up a barricade to keep players from walking off the cliff). In this way, the game industry has been at the cutting edge of research involving collecting users' information to understand and evaluate behaviors while playing a game (or using an application). For learning games, log data can provide a detailed record of learners' activities within the context of the game, and insight into the learning process. As such, the structure, ease of data collection on student learning, and the immediacy of feedback that is provided to learners can make games an ideal medium for assessing learning.

Challenges for Game-Based Assessment

Although there is a long history assessment in games, there is still limited use of video games as tools for assessment. In part, this is due to the need for more systematic research on how best to design in-game assessments that adequately estimate learners' knowledge and skills. However, there are still a number of broader challenges that need to be addressed stemming from the gaming context, as well as how games may be perceived. Specifically, challenges to the adoption of game-based assessment include the following:

1. *Lack of general acceptance of games as assessment tools.* Even with more open-ended areas of learning, there is still some resistance to the idea that something involving "play" can be a serious tool for learning. It has taken a while – and considerable empirical evidence –

for there to be a general acceptance that games can be effective tools to support student learning. For assessment, which has the potential to be even more "high stakes," the skepticism can be even greater that games can be useful tools for assessment. Overcoming this issue will involve ensuring that any game-based assessment is grounded in a robust theory of assessment and providing empirical support for the efficacy of game-based assessment.

2. *Test theoretical requirements.* Most games are designed in ways that the player advances a story by progressively solving problems that build on one another, where later challenges may require knowledge acquired in earlier parts of the game. This is in conflict with classical test theory, which assumes independence of test items, and often also with item response theory, which assumes local independence of items. This needs to be taken into consideration when designing games for assessment (Mislevy, Behrens, Dicerbo, Frezzo, & West, 2012).

3. *Possible extraneous cognitive load created by the gaming context.* Traditional assessments will often just ask children to say or write the information that is being assessed – or perhaps use the information in a simple way (e.g., "In the triangle presented below, solve the missing angle"). Alternative modes of assessment often ask for a more in-depth use of the content of what has been taught, for example, through solving of complex problems. Although this approach can tap into a "deeper" understanding, it also has the

potential to underestimate students' knowledge, because a small mistake early on in the solution can derail the entire process. Additionally, there is a possibility that the way the problem is presented (e.g., to provide complexity and/or authenticity) can add to complexity in ways that are not germane to the problem (i.e., they can add to the "extraneous cognitive load" in the assessment). This is also a potential issue when students are asked to apply knowledge to solve problems within the context of a game – in addition to the knowledge that is being assessed, additional knowledge or gaming skills may be required. Avoiding this particular issue involves paying close attention to the game mechanic being used for assessment to ensure that it is not adding unnecessary cognitive complexity.

4. *Games are meant for playing.* One of the first things experienced gamers will do when playing a new game is test the limits of the game. They will see how far off the path they can drive, or how they can "blow up" the system. Even if the game is intended as a form of assessment, learners may intentionally not solve a problem in the most efficient way in order to explore or play. Conversely, players may find a novel way to solve an in-game problem that does not require them to use the knowledge that is being assessed (i.e., they may find ways to "game" the system). There is a long tradition in games of both intentionally failing in order to test the limits of the system (e.g., finding out "how many hits will it take before my character dies?") as well as one of "cheats and hacks" (i.e., finding

ways to “win” that the programmers did not originally intend). Good design and playtest can help reduce the likelihood of either intentionally failing or “hacks.” It is also important to examine game logs to ensure that there were no unintended activities in the game.

In spite of these possible difficulties, the potential of game-based assessment is still great. To do it well, however, requires a grounding in established assessment theory, careful design of the in-game activities being used for assessment, and reviewing and evaluating in-game activities and assessments to determine their validity.

Foundations of Game-Based Assessment

A useful first step in understanding how to create reliable and effective assessment in video games is to examine how assessment has been conducted in other, related, digital systems. Thelwall (2000) argues that computerized assessments have passed through four distinct technological phases. The first generation of computerized testing began with the administration of conventional tests in a computerized format. The second generation included features that supported adaptive processes, attempting to tailor difficulty, content, or timing features of the subsequent item on the basis of examinees’ responses. During the third generation, advances in technology were influenced by item-response theory, resulting in adaptive measurement, including an automatic calibrated measurement system that continuously and unobtrusively estimated dynamic changes in the student’s achievement trajectory and profile. The fourth generation built on previous achievements in the adaptive computerized testing field by incorporating intelligent measurement, interpreting users' profiles, and providing advice to learners and teachers based on performance, which in turn is based on knowledge and inferencing procedures. Arguably, the body of research in which these advances in digital assessment have been most well developed and that has implication for game-based assessment is in the area of intelligent tutoring systems.

Assessment in Intelligent Tutoring Systems (ACT-R)

Research on assessments within the context of computer applications began decades ago, based on early information-processing models of human cognition (e.g., Koedinger, Anderson, Hadley & Mark, 1997). The Adaptive Control of Thought – Rational (ACT-R) theory was developed as a model of human thought that attempted to simulate how complex cognitive processes could arise through an interaction of more basic procedural and declarative knowledge (Anderson, Boyle, Corbett, & Lewis, 1990; Anderson, 1996). According to the ACT-R theory, human learning and cognition can be successfully modeled using computer language containing a set of procedural rules, composed of simple "if" and "then" statements. The ACT-R model is a network model that distinguishes between procedural knowledge, which involves a set of production rules, and declarative knowledge, which involves a database containing many units forming chunks of information (Miller, 1956; Servan-Schreiber, 1991; Anderson, 1996). The original ACT model followed a computationally plausible model of learning and memory, based on the theory that human memory is associative (Anderson & Bower, 1974). A major premise of the original ACT and revised ACT-R model is that all knowledge can be broken down into units of information, and a set of rules dictates how those units relate to one another. Unlike its predecessor, the ACT-R model accounts for the adaptive nature of learning by incorporating statistical structures better suited to explain the adaptive performance necessary to more accurately model human learning and memory processes.

In addition to modeling actual student learning and cognition, the ACT-R model was used to inform adaptive, or "intelligent," tutoring systems that would respond to student responses within the system. The ACT-R model was originally tested on knowledge acquisition related to problem-solving skills involved in mathematical or spatial reasoning (Anderson, Corbett, Koedinger, & Pelletier, 1995). In addition to predicting knowledge-acquisition processes, the model accounts for memory retrieval, or "knowledge deployment," processes and uses a method known as rational analysis (Anderson et al., 1995). According to the rational analysis framework, the availability of learned information can be predicted by the odds of it being used in a certain context and, therefore, depends on a set of conditional probabilities (Anderson, Boyle, Corbett, & Lewis, 1990; Anderson et al., 1995). The likelihood of correctly remembering some piece of information can be modeled using Bayesian inference to calculate the odds of the information being remembered given certain context and task-specific a priori probabilities. The ACT-R model predicts that human cognition tracks the overall usefulness of knowledge and assesses whether to apply the knowledge given a certain context. The likelihood of correctly retrieving information is therefore predicted by the effects of contextual priming. Based on the premises of this rational analysis framework proposed by Anderson and colleagues, a game environment context may increase the likelihood of knowledge being retrieved and applied correctly.

Research with these early cognitive tutoring programs indicates that the success of tutoring programs is largely due to the construction of separate models used to assess learning within the context in which knowledge is meant to be applied (e.g., Corbett, Anderson & O'Brien, 1995). Corbett and Anderson (1995) describe how the ACT Programming Tutor, intended to teach coding, constructs an *ideal student model* of knowledge in the domain (i.e., a "complete, executable model of procedural knowledge," p. 256). Students' actions are interpreted in light of this idealized model (a process termed "model tracing"), and if the actions indicate an understanding that falls short of the ideal model, then the system intervenes to support the students' learning. This discrepancy between the students' performance and the hypothesized learning model reveals the skills that the learner has mastery over and those that they have yet to develop.

This rational analysis framework proposed by Anderson and colleagues has been effectively used in tutoring specific academic skills; it applies well to more general game contexts for learning. Many computer games have a prespecified set of goals that the player must accomplish before a game-play session is considered successful. In most educational games, the achievement of these goals is evidence of the acquisition of a specific skill. When a player does not achieve the desired goals, much like the cognitive tutoring programs, the game allows them to continue by replaying segments that caused a discrepancy between their performance and mastery (i.e., the learning model). Given digital computer games' unobtrusive means of capturing students' performance within a contextually enriched learning environment, and considering earlier efforts (e.g., ACT-R tutoring programs), games as mediums for learning assessment seem to be a natural progression to create adaptive learning systems.

Theoretical Methods of Game-Based Assessment: Evidence-Centered Design

The approach to assessment within intelligent tutoring systems such as ACT-R provides some insight into how best to undertake game-based assessment. Further insight can be provided by considering appropriate models of assessment, one of the most promising of which is *evidence-centered design* (Mislevy & Haertel, 2006). ECD is a framework for developing assessment (Mislevy, Steinberg, & Almond, 2002) that asks two key questions: What knowledge, skills, etc. should be assessed, and what behaviors provide evidence of the knowledge, skills, etc.? As an approach for developing assessments, the ECD framework consists of interrelated components that describe the process of creating a conceptual assessment framework incorporating abstract knowledge aspects of the domain being assessed (i.e., *learning model*), paradigms for gathering information about domain proficiency (i.e., *performance model*), and the operational assessment whereby an instructor or administrator provides information and expectations necessary for completing the assessment as well as summary feedback to reciprocally improve the learner's future thinking and learning processes.

According to this framework, measuring proficiency involves consideration of at least three distinct paradigms that relate to standards of proficiency, evidence of proficiency, and the assessment task itself. Proficiency paradigms contain claims and aspects of proficiency; evidence paradigms consist of rubrics and means to identify evidence of proficiency in student work; task paradigms describe how students produce work relevant to the domain proficiency in question. The ECD framework promoted assessment validity and generalizability to measure aspects of learning, regardless of the content domain. In this regard, the ECD framework was based on an evidentiary perspective where criteria for proficiency is viewed as an argument and students' proficiency is determined by a body of logic-based evidence. In contrast to more traditional standardized forms of assessment, the ECD framework acknowledges the role of the assessment context and additional interrelated cognitive processes once considered peripheral to the proficiency domain. More recently, the *evidence-centered game design* (ECgD), a modified ECD framework for a game-based learning environment, has been developed (Mislevy, Orange, Bauer, von Davier, Hao, Corrigan, et al., 2014). Similar to the original ECD framework, ECgD defines targeted real-world competencies, aligns game-world to the real-world competencies, unobtrusively integrates formative feedback systems into games, and engages the learner in iterative design processes games with embedded assessment that support meaningful learning.

Within ECgD, there are four main components: (1) definition of real-world target competencies, (2) alignment of game-world competencies with those in the real-world context, (3) integration of unobtrusive formative feedback into the game, and (4) engagement in an iterative process to further develop engaging games with embedded assessment to deep learning.

Considering that the ECD framework effectively measures student proficiency, integrating assessments into game-based contexts seems sensible. The ECD framework acknowledges that mental and cognitive process of learning, students' activity, and observed performance are distinct aspects of the assessment framework. Digital computer games' affordance to monitor learners' activity enables such systems to collect evidence of domain proficiency on a continuous basis, allowing educators to better understand learner proficiency through documentation of the assessment context as well as learners' knowledge construction and application processes as documented through their actions and decisions within the game context.

Sources of Evidence for Assessment

When using ECD (or other assessment models) to guide game-based assessment, two fundamentally different approaches can be used regarding the activities that learners engage in to provide evidence for the assessment. One approach is to build activities into the game that are intended to evoke actions by the players that are to be used for assessment, and the other approach is to consider activities in the game after the fact, to look for evidence in game log files of the knowledge, skills, etc. that are being assessed. Both approaches have strengths and limitations, but either can be an effective source of assessment data.

Building Assessment into the Game: Assessment Mechanics. The term "game mechanics" refers to the specific actions, behaviors, and control mechanisms available to a player within the game (Hunicke, LeBlac, & Zubek, 2004). Building on the concept of game mechanics, Plass, Homer, Kinzer, Frye, and Perlin (2011) identify three discrete mechanics that impact the efficacy of educational games: *game mechanics*, *learning mechanics*, and *assessment mechanics*. Game mechanics has the same meaning as it does in nonlearning games; however, learning and assessment mechanics are considered "meta-mechanics" – they describe activities within the game designed with the intention of either supporting student learning or assessing the student (or both). A single learning mechanic or assessment mechanic may be instantiated as several different game mechanics, depending on game genre, platform, context, and users.

In an educational game, a learning mechanic may be a single activity or a set of coordinated activities that form the essential learning activities. Learning mechanics include high-level cognitive activities that describe how learning is presented to students, much like the task model in ECD. Depending on the educational content and type of game, the specific learning mechanic may vary. Nevertheless, it should always be grounded in the learning sciences and reflect activities that support student learning. For example, a learning mechanic might be that the player needs to apply specific math rules to solve a problem. This describes the learning function, but not the specific action within the game (i.e., the game mechanic). How this learning mechanic will be instantiated as a game mechanic could vary, similar to the presentation model in ECD. In a text-based game, the player may have to type out a response to a question in the game that requires using the math rule. If the game uses an Angry Birds–type "slingshot" game mechanic, then the player may need to fling the correct rule to the proper spot in a puzzle to solve the problem. In both cases, it is the same learning mechanic (use a rule to solve a math problem), but uses a different game mechanic (type a response versus "fling" a response).

Similarly, assessment mechanics involve an in-game activity or coordinated set of activities that is used to have players demonstrate knowledge or skills, similar to the task model in ECD. In the same way that the design of effective learning mechanics should be grounded in the learning sciences, the design of effective assessment mechanics should be grounded in theories of assessment, such as ECD or ECgD (Mislevy et al., 2014). Within the ECgD framework, the competency model assumes criteria indicative of certain competencies that are considered unobservable latent variables to be assessed. Meanwhile, the evidence model may collect and analyze behavioral evidence that is essentially observable variables to support users' relative understanding of the content or mastery of the skill. The evidence model also quantifies observable variables by establishing scoring systems to align evidence alongside claims statistically. Within this framework, designing assessment mechanics entails deciding what in-game activities can be created to provide data for the evidence model. Assessment mechanics may require the learner to apply rules to solve problems, to arrange items in time or space for solving problems, or to select items that are contiguous with either time or space (Plass, Homer, et al., 2013). For example, Light Lanes, a game developed by the NYU CREATE lab to teach about reflection and refraction and to promote systems thinking, asks users to direct a beam of light emitted from a fixed source into a vessel using reflectors that can be repositioned on a two-dimensional playing field. For this game, the positioning of the reflectors in space and the number of moves required to do so provide an opportune assessment mechanic. Other games may afford additional measures of performance (Reese, Seward, Tabachnick, Hitt, Harrison, & Mcfarland, 2012). For example, multiple scores may measure different performance areas such as total number of problems solved correctly, timed reports (Reese et al., 2012), or more complex performance indicators such as an "efficiency" score derived from a combined score of the proportion of items correct and the average response time (e.g., Homer, Plass, Raffaele, Ober, & Ali, 2018). Assessment mechanics in games may also serve as diagnostic instruments for complex cognitive and neurological disorders, such as dyslexia (Kyle, Kujala, Richardson, Lyytinen, & Goswami, 2013), dyscalculia (Wilson, Revkin, Cohen, Cohen, & Dehaene, 2006), and attention deficit/hyperactivity disorder (Rizzo, Buckwalter, Bowerly, Van Der Zaag, Humphrey, Neumann, et al., 2000).

When designing an assessment mechanic, a number of issues must be considered to assure its validity. Care must be taken to ensure that it does not depend too heavily on other factors that may present a confound to the assessment. For example, a game-based math assessment could be embedded into a baseball simulation that might then require a knowledge of baseball. If the student fails the assessment, it could be due to either not knowing the math or not knowing enough about baseball to understand the demands of the assessment. Similarly, depending on the gaming mechanic used, a player may fail the assessment because of a lack of gaming skills, rather than a lack of knowledge about the content being taught. For example, in one popular math game, learners are required to tilt their tablets to direct a ball to the correct answer. While this game mechanic is fun for the players, it is a poor assessment mechanic because failure on this task could be due to either a misconception in math or a lack of hand-eye coordination.

Conversely, it is also important to ensure that the task being used as an assessment cannot be passed in a way that does not require use of the knowledge of skill being assessed (e.g., through a "hack" or "cheat"). An example of this is reported in Shute, Venture, and Kim (2013), who were studying Newton's Playground, a video game that teaches physics to students. In the game, students must guide the path of a balloon by drawing simple machines (e.g., ramps, levers, pendulums). Shute et al. (2013) describe how they found in initial playtesting that students were initially able to pass some tasks just by drawing and stacking many small objects rather than using the simple machines that were supposed to be used for the tasks. The authors eliminated this problem by imposing a limit on the number of items students could draw to solve any one problem.

Another issue to consider is that mechanics that have been designed to optimize learning may not be ideal for assessment. For example, the Alien Game has been shown to be effective for developing high school students' executive functions (Parong, Mayer, Fiorella, McNamara, Homer & Plass, 2017; Homer et al., 2018). However, because the game is adaptive, there is too much variability in performance demands within the game for any simple metrics in the game to serve as an effective assessment mechanic.

A final issue to consider when developing assessment mechanics is the role of emotion. Good games are intended to elicit emotions from players, and in part, it is this emotional engagement that allows games to be excellent tools for learning and assessment (Plass et al., 2015). However, too much emotional arousal in an educational game can result in excessive cognitive load and interfere with learning and assessment (Fraser, Ma, Teteris, Baxter, Wright, & McLaughlin, 2012). Assessments within games, then, should be emotionally engaging but not excessively emotionally arousing.

Assessing In-Game Activities after the Fact: Computer Log Data. A second source of data for in-game assessment comes from log data (Shute, 2011; Plass, Homer, et al., 2013; All, Nunez Castellar, & Van Looy, 2016). In this case, the specific assessment activities may not be predefined. Instead, to use the ECgD framework, data from the game are examined statistically to create an evidence model for certain target competencies (Mislevy et al., 2014). The best examples of this approach come from work by Shute and her colleagues in the "stealth assessment" approach (Shute, 2011; Shute & Ventura, 2013; Shute & Sun, in press). This method uses Bayesian network analysis (de Klerk, Veldkamp, & Eggen, 2015, 2016) in order establish conditional relationships among in-game indicators and the competency variables. Each of these conditional relationships is determined by a set of statistical probabilities assigned based on the user's previous course of action. The use of advanced statistical methods for analyzing computer log data may allow researchers to draw more accurate inferences of the user's competency model based on the information available about the user's performance model. The stealth-assessment approach has been used to successfully evaluated learner's creativity, problem-solving, spatial skills, and persistence (Shute, 2011; Buelow, Okdie, & Cooper, 2015).

Analyses of log assessment data can also be used to enhance the effectiveness of an educational intervention. For example, Baker, Clarke-Midura, and Ocumpaugh (2016) studied log file data collected from 2,000 middle school students as they explored a virtual world intended to support science education. The authors then examined models of behavior in the virtual world that predicted science inquiry and achievement. The virtual environment featured many characteristics of an exploration game with personalized avatars and opportunities for goal-setting and served as a virtual performance assessment, extracting information about students' sequence of actions and response times to build a probabilistic model of student performance in relation to target competencies indicative of science inquiry. This probabilistic model was then applied to a new scenario or "virtual world" to test users' skills of scientific inquiry based on parameters that had been operationalized in the previous scenario. The results indicated that the probabilistic model, when applied to a new scenario that featured similar structural components of the original one, reliably predicted student performance in the new scenario. Furthermore, the probabilistic model of performance also identified students unable to demonstrate scientific inquiry skills, allowing for early intervention. The model was capable of identifying learners with poor self-regulation: those demonstrating off-task behavior, inadequate or extended amount of time spent on a certain task, or behaviors indicative of frustration, boredom, or lack of perseverance.

While computer log data use provides the user, teacher, and researcher with invaluable information, such an abundance of data may be challenging to read and interpret. Solutions to this problem include generating log files that record only information of interest (see Shute, Wang, Greiff, Zhao, & Moore, 2016) or developing a generic log file structure applicable to different games to manage the more tedious aspects of data storage and extraction (see Hao, Smith, Mislevy, von Davier, & Bauer, 2016).

Validation of Game-Based Assessments

As with any new assessment tool, game-based assessments need to be validated. Three main approaches for such a validation include: (1) comparing game-based assessments to established assessment tools given outside the context of the game, (2) having experts evaluate in-game activities and comparing expert evaluation with the in-game assessment, or (3) examining how well the game-based assessments predict some sort of future learning performance or outcome in the domain.

The most common approach is to give external measures of the knowledge, skills, etc. that are being assessed within the game. For example, the game Factor Reactor was created to help develop math fluency in middle school children (see Plass, O'Keefe, et al., 2013). The game presents users with a center number surrounded by two rings of other numbers. The objective is to transform the center number into one of the outer "goal" numbers by adding, subtracting, multiplying, or dividing it by one of the numbers in the inner ring. Players are rewarded for both speed and using the fewest possible number of steps to transform the target number. In a study examining the effects of different play conditions (individual, competitive, or collaborative), Plass, O'Keefe, et al. (2013) used the Woodcock–Johnson III Math Fluency subtest (McGrew & Woodcock, 2001) as an external measure of math fluency and found that a number of game metrics, including levels completed, were correlated with students' scores in the Woodcock–Johnson test, suggesting that the game could be a valid assessment tool (Plass, Homer et al., 2013; Plass, O'Keefe, et al., 2013).

Another example of using external validation of game-based assessment comes from a study by Shute et al. (2016), who sought to design, develop, and validate a game-based assessment to measure problem-solving abilities with a group of middle school students playing a custom-designed game, Use Your Brainz. The students played the game on a mobile tablet for three hours over the course of three consecutive days. On the fourth day, they participated in a series of post-tests to measure far transfer as a result of game play. The game Use Your Brainz, which was closely modeled on the structure of the popular game Plants vs. Zombies 2, had been previously used as an instrument for inquiry into the efficacy of game-based stealth assessment (Shute, Moore, & Wang, 2015). In designing the game, the researchers constructed a competency model based on previous research on problem-solving. Aspects of problem-solving included analyzing game constraints, planning a course of action toward a solution, using resources efficiently, and monitoring progress along the way toward a solution. In addition, a set of actions was identified that indicated either the acquisition or application of a certain rule by the learner, necessary for achieving a solution. Bayesian network analysis was used to evaluate the progress of learners while playing the game to construct an evidence model with respect to the desired competency model. The results from the in-game measures were then compared with two external measures, MicroDYN (Wüstenberg, Greiff, & Funke, 2012) and Raven's Progressive Matrices (Raven, 1941, 2000). A multiple regression model revealed that both external measures predicted some of the variability within game measures, with the MicroDYN subscore for knowledge application significantly predicting various within-game measures such as planning, tool usage, and evaluation progress. This research suggests that the greater complexity that games afford assessment mechanics may make them ideal environments to test abstract constructs that require cognitive flexibility and creative problem-solving.

Practical Implications for Education

Assessment in the context of games holds much promise for improving testing practices, especially for formative assessments. Game-based assessment may promote mastery learning with students while simultaneously allowing teachers and researchers to close the gap between the desired competency model and the individual student's performance model by collecting detailed information about their progress. Furthermore, with quality design, game-based assessments may promote a more authentic form of knowledge construction, whereby the learner can acquire practical knowledge that utilizes skills such as problem-solving (Kiili, 2007) and spatial reasoning (De Lisi & Wolford, 2002). These authentic forms of learning may ultimately promote knowledge transfer, while stimulating long-term retention and retrieval processes through sustained attention and engagement within the game learning environment.

Limitations of Game-Based Assessments

In order for game-based assessments to be practical and achieve the desired aims of assessing students within a dynamic learning environment, they must be well designed and cater to learners and researchers, teachers, or administrators. Data points must be relevant and easily interpretable (Leighton & Chu, 2016). Unfortunately, the cost of developing of a high-quality game-based assessment system can be difficult to justify. Additional concerns revolve around the issue of fairness, particularly as testing is involved. Most games afford users a context to learn and explore content; however, if a learner is unfamiliar with the context or setting of the game, it may place them at an unfair disadvantage for learning the material (Kim & Shute, 2015). Conversely, learners who are avid gamers may have an advantage within the context of the game that does not serve them well in a real-world setting.

An open question concerns the usefulness of games as standardized assessments. A fully developed game may introduce too many confounds, as discussed above, but it seems clear that current approaches to standardized assessments would benefit from insights resulting from game-based learning and assessment, such as their affordance to provide meaningful contexts for performance, to incorporate emotional design considerations, and to motivate the learner or test taker (Plass et al., 2016).

Future Research

While the findings described above are ultimately promising, further research is necessary to understand the full implications and possibilities for game-based assessment. Future research on game-based assessment should consider three basic questions as outlined by Mayer (2015) with respect to educational games in general. The first question seeks to address the value-added nature of "game as embedded assessment tool" and attempts to identify the underlying benefits of using games for assessment purposes, such as optimizing instructional time, instantaneous formative feedback, and facilitating engagement. The second question addresses the cognitive consequences of using games for assessment. For example, is a student in fact learning the content that is relevant for long-term achievement, or is the student merely demonstrating optimal performance in the game because of an understanding of the game design? The third question addresses the issue of media comparison and whether digital or traditional forms of media are more suitable for certain content. For example, would students benefit more from playing the game in a digital or traditional context? Also, which format is most likely to lead to near or far transfer?

While these three broad questions provoke long-term consideration of the use of games as assessment tools, an understanding of games as assessment tools may benefit from integration methods to detect learners' cognitive, affective, and motivational responses. The inclusion of emotion recognition along with assessment may allow future digital technologies to mirror human interactions and positively influence the learner's performance. Referred to as affective computing (Picard & Picard, 1997), such technologies could detect and respond accordingly to human emotions that may serve as effective means to mitigate negative affect (e.g., frustration and anxiety induced during the learning and testing process) and optimize long-term learning by adapting the game environment to provide a context that induces positive emotions such as confidence and fascination (Novak & Johnson, 2012). Positive affect is associated with improved long-term memory outcomes that support working memory, storage, and retrieval processes (Erez & Isen, 2002), often viewed as indicators of actual learning (Chen & Wang, 2011). In addition to research on affective computing, augmented and virtual reality are rapidly becoming more accessible to a broader set of users and may serve as means to further develop models for authentic and engaging assessment.

Conclusions

In this chapter, we examined the ways in which digital games can be used to authentically evaluate learners' knowledge and skills. Challenges to the use of game-based assessment include lack of general acceptance of games as assessment tools, potential for extraneous cognitive load caused by the gaming environment, test theoretical concerns related to item independence, and a culture of exploration and "cheats/hacks" in games. To overcome these challenges, we have argued that game-based assessments need to be grounded in existing assessment practice and theory (e.g., ECD/ECgD), undergo thorough playtesting (including evaluation of the activities being used to assess learners' knowledge, i.e., the assessment mechanics), and be validated through external evaluations and examination of game log data. The integration of assessment into the context of games offers the promise of constructing a high-quality dynamic system that is engaging and adaptive to the learner while assessing student knowledge in order to enhance student learning. Though game-based assessments hold much promise with future implications for education, more research is needed to fully address limitations and questions regarding its practical usage.

chapter-references

References

All, A., Castellar, E. P. N., & Van Looy, J. (2016). Assessing the effectiveness of digital game-based learning: Best practices. *Computers & Education*, 92–*93*, 90–103.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365.

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition*, *2*(3), 406–412.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial intelligence*, *42*(1), 7–49.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207.

Azevedo, R., Cromley, J. G., Moos, D. C., Greene, J. A., & Winters, F. I. (2011). Adaptive content and process scaffolding: A key to facilitating students' self-regulated learning with hypermedia. *Psychological Testing and Assessment Modeling*, *53*(1), 106–140.

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

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* (formerly Journal of Personnel Evaluation in Education), *21*(1), 5.

Buelow, M. T., Okdie, B. M., & Cooper, A. B. (2015). The influence of video games on executive functions in college students. *Computers in Human Behavior*, *45*, 228–234.

Chen, C. M., & Wang, H. P. (2011). Using emotion recognition technology to assess the effects of different multimedia materials on learning emotion and performance. *Library & Information Science Research*, *33*(3), 244–255.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. *Cognitively Diagnostic Assessment*, 19–41.

Csapó, B., Lörincz, A., & Molnár, G. (2012). Innovative assessment technologies in educational games designed for young students. In *Assessment in game-based learning* (pp. 235–254). New York: Springer.

De Lisi, R., & Wolford, J. L. (2002). Improving children's mental rotation accuracy with computer game playing. *Journal of Genetic Psychology*, *163*(3), 272–282.

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

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2016). A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in Bayesian network. *Computers in Human Behavior*, *60*, 264–279.

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science*, *238*, 1556–1557.

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior*, *26*(5), 1024–1033.

Fraser, K., Ma, I., Teteris, E., Baxter, H., Wright, B., & McLaughlin, K. (2012). Emotion, cognitive load and learning outcomes during simulation training. *Medical Education*, *46*(11), 1055–1062.

Erez, A., & Isen, A. M. (2002). The influence of positive affect on the components of expectancy motivation. *Journal of Applied Psychology*, *87*(6), 1055.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Macmillan.

Gee, J. P., & Shaffer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. *Phi Delta Kappa International EDge*, *6*(1), 3–19.

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). Taming log files from game/simulation-based assessments: Data models and data analysis tools. Research Report ETS RR-16-10. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12096/epdf.

Homer, B. D., Plass, J. L., Raffaele, C., Ober, T. M., & Ali, A. (2018). Improving high school students' executive functions through digital game play. *Computers & Education*, *117*, 50–58.

Hoffman, B., & Nadelson, L. (2010). Motivational engagement and video gaming: A mixed methods study. *Educational Technology Research and Development*, *58*(3), 245–270.

Hunicke, R., LeBlanc, M., & Zubek, R. (2004, July). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI.*

Kiili, K. (2007). Foundation for problem-based gaming. *British Journal of Educational Technology*, *38*(3), 394–404.

Kim, B. (2015). Game mechanics, dynamics, and aesthetics. *Library Technology Reports*, *51*(2), 17.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, *87*, 340–356.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. In *Proceedings of the 7th World Conference on Artificial Intelligence in Education*. Charlottesville, VA: Association for the Advancement of Computing in Education.

Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., & Goswami, U. (2013). Assessing the effectiveness of two theoretically motivated computer-assisted reading interventions in the United Kingdom: GG Rime and GG Phoneme. *Reading Research Quarterly*, *48*(1), 61–76.

Leighton, J. P., & Chu, M. W. (2016). First among equals: Hybridization of cognitive diagnostic assessment and evidence-centered game design. *International Journal of Testing*, *16*(2), 164–180.

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, *13*(2), 125–139.

Mayer, R. E. (2015). On the need for research evidence to guide the design of computer games for learning. *Educational Psychologist*, *50*(4), 349–353.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock–Johnson III*. Itasca, IL: Riverside.

Meyer, B., & Sørensen, B. H. (2009). Designing serious games for computer assisted language learning: A framework for development and analysis. In M. Kankaanranta & P. Neittaanmäki (Eds.), *Design and use of serious games* (pp. 69–82). Dordrecht: Springer.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In *Assessment in game-based learning* (pp. 59–81). New York: Springer.

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., Orange, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., … & John, M. (2014). Psychometric considerations in game-based assessment. White paper. Retrieved from www.ets.org/research/policy_research_reports/publications/white_paper/2014/jrrx.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496.

Novak, E., & Johnson, T. E. (2012). Assessment of student’s emotions in game-based learning. In *Assessment in game-based learning* (pp. 379–399). New York: Springer.

Parong, J., Mayer, R. E., Fiorella, L., MacNamara, A., Homer, B. D., & Plass, J. L. (2017). Learning executive function skills by playing focused video games. *Contemporary Educational Psychology*, *51*, 141–151.

Piaget, J. (1962). *Play, dreams and imitation in childhood*. New York: W. W. Norton.

Picard, R. W., & Picard, R. (1997). *Affective computing* (Vol. 252). Cambridge: MIT Press.

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*, 258–283.

Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., … & Perlin, K. (2013). Metrics in simulations and games for learning. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics* (pp. 697–729). London: Springer.

Plass, J. L., Homer, B. D., Kinzer, C. K., Frye, J., & Perlin, K. (2011). Learning mechanics and assessment mechanics for games for learning. G4LI White Paper, 1.

Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., & Perlin, K. (2013). The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology*, *105*(4), 1050–1066.

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

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137–150.

Ravitch, D. (2016). *The death and life of the great American school system: How testing and choice are undermining education*. Basic Books.

Reese, D. D., Seward, R. J., Tabachnick, B. G., Hitt, B. A., Harrison, A., & Mcfarland, L. (2012). Timed report measures learning: Game-based embedded assessment. In *Assessment in game-based learning* (pp. 145–172). New York: Springer.

Rizzo, A. A., Buckwalter, J. G., Bowerly, T., Van Der Zaag, C., Humphrey, L., Neumann, U., ... & Sisemore, D. (2000). The virtual classroom: A virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior*, *3*(3), 483–499.

Servan-Schreiber, E. (1991). The competitive chunking theory: Models of perception, learning, and memory. Doctoral dissertation, Carnegie-Mellon University.

Shute, V., & Sun, C. (in press). Game-based assessment: What it is and does it work? In J. Plass, R. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning*. MIT Press.

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

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., Moore, G. R., & Wang, L. (2015). Measuring problem solving skills in Plants vs. Zombies 2. *International Educational Data Mining Society: Proceedings of the 8th International Conference on Educational Data Mining* (pp. 428–431).

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.

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

Thelwall, M. (2000). Computer-based assessment: A versatile educational tool. *Computers & Education*, *34*(1), 37–49.

US Department of Education (2017). Reimagining the role of technology in education: 2017 National Education Technology Plan Update. Retrieved from https://tech.ed.gov/files/2017/01/NETP17.pdf.

Wilson, A. J., Revkin, S. K., Cohen, D., Cohen, L., & Dehaene, S. (2006). An open trial assessment of “The Number Race,” an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, *2*(1), 20.

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