Stealth Assessment in Digital Games
Valerie Shute, Matthew Ventura, and Diego Zapata-Rivera

Education in the 21st Century

Preparing our kids to succeed in the 21st century requires fresh thinking on how to cultivate a new set of educationally-valuable competencies.\(^1\) In addition to identifying and modeling new competencies, there’s an associated need to design and develop valid assessments of them and support their development. The reason that new thinking on educational competencies is needed is because in the first half of the 20th century, a person who acquired basic reading, writing, and math skills was considered to be sufficiently literate to enter the work force (Kliebard, 1987). The goal back then was to prepare young people as service workers because 90% of the students were not expected to seek or hold professional careers (see Shute, 2007). But with the emergence of the Internet, the world has become more interconnected, effectively smaller, and more complex than before.

Developed countries now rely on their knowledge workers to deal with an array of complex problems, many with global ramifications (e.g., climate change, alternative energy sources). Knowledge workers, as well as informed citizens in general, when confronted by such problems, need to be able to think systemically, creatively, and critically (e.g., Shute & Torres, 2011; Walberg, & Stariha, 1992). These skills are a few of what many educators are calling 21st century (or complex) competencies.

\(^1\)We use the term "competency" broadly to cover a range of attributes such as knowledge, skills, understanding, dispositions, and values.
Except in rare instances, our current education system neither teaches nor assesses these new competencies despite a growing body of research showing that competencies, such as persistence, creativity, self-efficacy, openness, and teamwork (to name a few) can substantially impact student academic achievement (Noftle & Robins, 2007; O’Connor & Paunonen, 2007; Poropat, 2009; Sternberg, 2006; Trapmann, Hell, Hirn, & Schuler, 2007). Furthermore, traditional assessments of content are often too simplified, abstract, and decontextualized to suit current education needs, and they also fail to assess what students actually can do with the knowledge and skills obtained inside and outside of school (Shute, 2009). However, well-designed digital games can provide meaningful assessment environments by providing students with scenarios that require the application of various competencies.

For example, consider role playing games (e.g., World of Warcraft). In these games, players must read lengthy and complex quest logs that tell them the goals in the game. Without comprehending these quest instructions, the players would not be able to know how to proceed and succeed in the game. This seemingly simple task in role playing games is, in fact, an authentic, situated assessment of reading comprehension. Without these situated and meaningful assessments, we cannot determine what students can actually do with the skills and knowledge obtained. Thus new, embedded, authentic types of assessment methods are needed to properly assess valued competencies.

Why use well-designed games as vehicles to assess and support learning? There are several reasons. First, because our schools have remained virtually unchanged for many decades while our world is changing rapidly, we are seeing a growing number of disengaged students. This disengagement increases the chances of students dropping out of school. For instance, high dropout rates, especially among Hispanic, Black, and Native American students, were described
as “The Silent Epidemic” in a 2006 research report for the Bill and Melinda Gates Foundation. According to this report, nearly one third of all public high school students dropout and the rate is higher for minority students. In the report, when 467 high school dropouts were asked why they left school, 47% of them simply responded, “The classes were not interesting.” We need to find ways (e.g., well-designed digital games and other immersive environments) to get our kids engaged, support their learning, and allow them to contribute fruitfully to society.

A second reason for using games as assessments is a pressing need for dynamic and ongoing measures of learning processes and outcomes. Interest in alternative forms of assessment is driven by dissatisfaction with and limitations of multiple-choice items. In the 1990s, interest in alternative forms of assessment increased with the popularization of what became known as authentic assessment. A number of researchers found that multiple-choice and other fixed-response formats substantially narrowed school curricula by emphasizing basic content knowledge and skills within subjects and not assessing higher-order thinking skills (e.g., Kellaghan & Madaus, 1991; Shepard, 1991). However, as Madaus and O’Dwyer (1999) argued, incorporating performance assessments into testing programs is difficult because they are less efficient, more difficult and disruptive to administer, and more time-consuming than multiple-choice testing programs. Consequently, multiple-choice has remained the dominant format in most K-12 assessments in our country. New performance assessments are needed that are valid, reliable, and automated in terms of scoring.

A third reason for using games as assessment vehicles is that many games typically require a player to apply various competencies to succeed in the game (e.g., creativity, problem solving, persistence, and collaboration). The competencies required to succeed in many games also happen to be the same ones that companies are looking for in today’s highly competitive
economy (Gee, Hull, & Lankshear, 1996). Moreover, games are a significant and ubiquitous part of young people’s lives. For instance, the Pew Internet and American Life Project surveyed 1,102 youth between the ages of 12 and 17. They reported that 97% of youth – both boys (99%) and girls (94%) – play some type of digital game (Lenhart et al., 2008). Additionally, Ito and her colleagues (2010) found that playing digital games with friends and family is a large and normal part of the daily lives of youth. They further observed that playing digital games is not solely for entertainment purposes. In fact, many youth participate in online discussion forums to share their knowledge and skills about a game with other players, or seek help on challenges when needed.

In addition to the arguments for using games as assessment devices, there is growing evidence of games supporting learning (e.g., Shute, Ventura, Kim, & Wang, in press; Wilson et al., 2009). However, we need to understand more precisely how and what kinds of knowledge and skills are being acquired (Vorderer & Bryant, 2006). Understanding the relationships between games and learning is complicated by the fact that we don’t want to disrupt players’ engagement levels during game play. Consequently, learning in games has historically been assessed indirectly and/or in a post hoc manner (Shute & Ke, in press). What’s needed instead is real-time assessment and support of learning based on the dynamic needs of players. We need to be able to experimentally ascertain the degree to which games can support learning, and how and why they achieve this objective.

This paper presents the theoretical foundations of and the research methodologies for designing, developing, and evaluating stealth assessments in digital games. Very generally, stealth assessments are assessments embedded within games to unobtrusively, accurately, and dynamically measure how players are progressing relative to targeted competencies (Shute, 2011; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Embedding assessments within games
provides a way to monitor a player’s current level on valued competencies, and then use that information as the basis for support, such as adjusting the difficulty level of challenges or providing timely feedback.

There are four main sections in this paper. First, we discuss problems with existing traditional assessments. Second, we review evidence relating to digital games and learning. Third, we define and then illustrate our stealth assessment approach with a set of assessments we are currently developing and embedding in a digital game (Crayon Physics Deluxe). The stealth assessments are intended to measure levels of creativity, persistence, and conceptual understanding of Newtonian physics during game play. Finally, we discuss future research and issues related to stealth assessment in education.

**Problems with Current Assessments**

Our country’s current approach to assessing students has a lot of room for improvement, at the classroom and high-stakes levels. This is especially true in terms of the lack of support that standardized, summative assessments provide for students learning new knowledge, skills, and dispositions that are important to succeed in today’s complex world. The current means of assessing students infrequently (e.g., at the end of a unit or school year, for grading and promotion purposes) can cause various unintended consequences, such as increasing the dropout rate given out-of-context and often irrelevant test-preparation teaching contexts that the current assessment system often promotes.

The goal of an ideal assessment policy/process should be to provide valid, reliable, and actionable information about students’ learning and growth that allows stakeholders (e.g., teachers, students, administrators, and parents) to utilize the information in meaningful ways.
Before describing particular problems associated with current assessment practices, we first provide a brief overview of assessment.

**Assessment, Writ Large**

People often confound the concepts of “measurement” and “assessment.” Whenever you need to measure something accurately, you use an appropriate tool to determine how tall, short, hot, cold, fast, or slow something is. We measure to obtain information (data), which may or may not be useful, depending on the accuracy of the tools we use, as well as our skill at using them. Measuring things like a person’s height, a room’s temperature, or a car’s speed is technically not an assessment, but rather the collection of information relative to an established standard (Shute, 2009).

**Educational measurement.** Educational measurement refers to the application of a measuring tool (or standard scale) to determine the degree to which educationally-valuable knowledge, skills, and other attributes have been, or are being acquired. It involves the collection and analysis of learner data. According to the National Council on Measurement in Education Web site, this includes theory, techniques, and instrumentation available for measurement of educationally-relevant human, institutional, and social characteristics. A test is education’s equivalent of a ruler, thermometer, or radar gun. But a test does not improve learning any more than a thermometer cures a fever; both are simply tools. Moreover, as Snow and Jones (2001) point out, tests, alone, cannot enhance educational outcomes. Rather, tests can guide improvement (given they are valid and reliable) if they motivate adjustments to the educational system (i.e., provide the basis for bolstering curricula, ensure support for struggling learners, guide professional development opportunities, and distribute limited resources fairly).
Again, we measure things in order to get information, which may be quantitative or qualitative. How we choose to use the data is a different matter. For instance, back in the early 1900s, students’ abilities and intelligence were extensively measured. However, this wasn’t done to help them learn better or otherwise to progress. Instead, the main purpose of testing was to track students into appropriate paths, with the understanding that their aptitudes were inherently fixed. A dominant belief during that period was that intelligence was part of a person’s genetic makeup thus testing was aimed specifically at efficiently assigning students into high, middle, or low educational tracks according to their supposedly innate mental abilities (Terman, 1916). In general, there was a fundamental shift to practical education going on in the country during the early 1900s, countering “wasted time” in schools and abandoning the classics as useless and inefficient for the masses (Shute, 2007). Early educational researchers and administrators inserted into the national educational discourse the metaphor of the school as a “factory” (Kliebard, 1987). This metaphor has persisted to this day.

**Assessment.** Assessment involves more than just measurement. In addition to systematically collecting and analyzing information (i.e., measurement), it also involves interpreting and acting on information about learners’ understanding and/or performance relative to educational goals. Measurement can be viewed as a subset of assessment.

As mentioned earlier, assessment information can be used by a variety of stakeholders and for a variety of purposes (e.g., to help improve learning outcomes, programs, and services; and also to establish accountability). Furthermore, there is an assortment of procedures associated with the different purposes. For example, if your goal was to enhance an individual’s learning, and you wanted to determine her progress toward an educational goal, you could: (a) administer a quiz; (b) view a portfolio of her work; (c) ask the student (or peers) to evaluate her
progress; (d) watch the person solve a complex task, (e) review her lab reports or journal entries, and so on.

In addition to having different purposes and procedures for obtaining information, assessments may also be differentially referenced or interpreted; for example, in relation to normative data or to a criterion. Norm-referenced interpretation compares learner data to that of other individuals or to a larger group, but can also involve comparisons to oneself (e.g., asking a person how she’s feeling and getting a, “Better than usual” response is a norm-reference interpretation). The purpose of norm-referenced interpretation is to establish what is typical or reasonable. On the other hand, criterion-referenced interpretation involves establishing what a person can or cannot do, or typically does or does not do—specifically in relation to a criterion. If the purpose of the assessment is to support personal learning, then criterion-referenced interpretation is required (for more, see Nitko, 1980).

This overview of assessment is intended to provide a foundation for the next section where we examine specific problems surrounding current assessment practices.

**Traditional Classroom Assessments Are Detached Events**

Current approaches to assessment are usually divorced from learning. That is, the typical educational cycle is: Teach. Stop. Administer test. Go loop (with new content). But consider the following metaphor representing an important shift that occurred in the world of retail outlets (from small businesses to supermarkets to department stores), suggested by Pellegrino, Chudowsky, and Glaser (2001, p. 284). No longer do these businesses have to close down once or twice a year to take inventory of their stock. Rather, with the advent of automated checkout and barcodes for all items, these businesses have access to a continuous stream of information that can be used to monitor inventory and the flow of items. Not only can a business
without interruption, but the information obtained is far richer than before, enabling stores to
monitor trends and aggregate the data into various kinds of summaries, as well as to support real-
time, just-in-time inventory management. Similarly, with new assessment technologies, schools
should no longer have to interrupt the normal instructional process at various times during the
year to administer external tests to students. Instead, assessment should be continual and
invisible to students, supporting real-time, just-in-time instruction (for more, see Shute, Levy,

**Traditional Classroom Assessments Rarely Influence Learning**

In addition, many of today’s classroom assessments fail to support deep learning or the
acquisition of complex competencies. Current classroom assessments are typically designed to
judge a student (or group of students) at a single point in time, without providing diagnostic
support to students or diagnostic information to teachers. Instead, assessments should be used to:
(a) support the learning process for students and teachers; (b) interpret information about
understanding and/or performance regarding educational goals (local to the curriculum, and
broader to the state or Common Core Standards); (c) provide formative, compared to summative
information (e.g., give useful feedback during the learning process rather than a single judgment
at the end); and (d) be responsive to what’s known about how people learn – generally and
developmentally.

To illustrate how a classroom assessment may be used to support learning, Shute, Hansen,
and Almond (2008) conducted a study to evaluate the efficacy of an assessment for learning
system named ACED (Adaptive Content with Evidence-based Diagnosis). They used an ECD
approach (Mislevy, Steinberg, & Almond, 2003) to create an adaptive, diagnostic assessment
system which also included instructional support in the form of elaborated feedback. The key
issue examined was whether the inclusion of the feedback into the system (a) impairs the quality of the assessment (relative to validity, reliability, and efficiency), and (b) does, in fact, enhance student learning. Results from a controlled evaluation testing 268 high-school students showed that the quality of the assessment was unimpaired by the provision of feedback. Moreover, students using the ACED system showed significantly greater learning of the content (geometric sequences) compared with a control group. These findings suggest that assessments in other settings (e.g., state-mandated tests) might be augmented to support student learning with instructional feedback without jeopardizing the primary purpose of the assessment.

**Traditional Assessment and Validity Issues**

Assessments are typically evaluated under two broad categories: reliability and validity. Reliability is the most basic requirement for an assessment and is concerned with the degree to which a test can consistently measure some attribute over similar conditions. In assessment, reliability is seen when a person scores really high on an algebra test at one point in time and then scores similarly on a comparable test 1-2 days later. In order to achieve high reliability, assessment tasks are simplified to independent pieces of evidence which can be modeled by existing measurement models.

An interesting issue is how far this simplification process can go without negatively influencing the validity of the test. That is, in order to remove any possible source of construct irrelevant variance and dependencies, tasks end up looking as decontextualized, discrete pieces of evidence. In the process of achieving high reliability, which is important for supporting high-stakes decision making, other aspects of the test are sacrificed (e.g., engagement and some aspects of validity).
Another aspect that traditional, standardized summative assessments emphasize is dealing with operational constraints (e.g., the need for gathering and scoring sufficient pieces of evidence within a limited administration time and budget). In fact many of the simplifications done could be explained by this issue and by the current state of measurement models that do not handle complex tasks interactions, the presence of feedback, and student learning during the test.

Validity, broadly, refers to the extent to which an assessment actually measures what it is intended to measure. Here are specific validity issues related to traditional assessment.

**Face validity.** Face validity states that an assessment should intuitively “appear” to measure what it is intended to measure. For example, reading some paragraphs on an uninteresting or irrelevant topic and answering multiple-choice questions about it is not the best measure for reading comprehension (i.e., it lacks good face validity). As suggested earlier, students need to be assessed in meaningful environments rather than filling in bubbles on a prepared form in response to decontextualized questions. Digital games can provide such meaningful environments by providing students with scenarios that require the application of various competencies, such as reading comprehension and problem solving.

**Predictive validity.** Predictive validity refers to an assessment predicting future behavior. Today’s large-scale, standardized assessments are lacking in this area. For example, a recent report from the College Board found the SAT only marginally predicted college success beyond high school GPA at around $r = .10$ (Korbin, Patterson, Shaw, Mattern, & Barbuti, 2008). This means the SAT scores contribute around 1% of unique prediction to college success after controlling for GPA information. However, a single construct like “grit” (i.e., persistence plus passion) has been shown to account for an average of 4% of the variance related to various academic outcomes (Duckworth, Peterson, Matthews, & Kelly, 2007). This does not necessarily
mean that tests like the SAT are redundant and unnecessary. In fact, some could argue that a test that can provide information similar to that of the GPA is useful taking into account problems such as grade inflation. These findings suggest a need for assessments that measure important new constructs and predict what kids can do beyond the test.

**Consequential validity.** Consequential validity refers to the effects of a particular assessment on society and policy decisions. One negative side effect of the No Child Left Behind (NCLB, 2002) initiative, with its heavy focus on accountability, has been teachers “teaching to the test” thereby reducing the face validity of such tests. That is, when teachers instruct content that is relevant to answering items on a test but not particularly relevant for solving real-world problems, this reduces student engagement in school which can lead to increased dropouts (Bridgeland, DiIulio, & Morison, 2006). Moreover, the lack of predictive validity of current assessments can lead to students not getting into college due to the inaccuracy or inappropriateness of the assessments. But the SAT and similar test scores are still being used as the main basis for college admission decisions which can lead to many students missing opportunities at fulfilling careers and lives, particularly disadvantaged youth. There is thus a need for new assessments that focus on relevant skills such as persistence and others, which could lead to positive (not negative) consequences for students.

To illustrate the contrast between traditional and new performance-based assessments, consider the attribute of conscientiousness. Conscientiousness can be broadly defined as the motivation to work hard despite challenging conditions, a disposition which has consistently been found to predict academic achievement from preschool, to high school, to the postsecondary level, and adulthood. Conscientiousness measures, like most dispositional measures, are primarily self-report (e.g., I work hard no matter how difficult the task; I
accomplish my work on time), a method of assessment that is riddled with problems. First, self-report measures are subject to “social desirability effects” that can lead to false reports about behavior, attitudes, and beliefs. Second, self-report measures can be easily coached. That is, test takers can be instructed with little effort how to respond “correctly” on a self-report measure. Good games, coupled with evidence-based assessment, show promise as a vehicle to dynamically measure conscientiousness and other important competencies more accurately than traditional approaches. These evidence-based assessments can record and score multiple behaviors and measurable artifacts in the game that pertain to particular competencies.

For instance, various actions a player takes within a well-designed game can inform conscientiousness—how long a person spends on a difficult problem (where longer = more persistent), the number of failures and re-tries before success, returning to a hard problem after skipping it, and so on. Each instance of these “conscientiousness indicators” would update the student model of this variable—thus would be up-to-date and available to view at any time. Additionally, good games provide a game play environment that can improve conscientiousness, because many problems require players to persevere despite failure and frustration. That is, many good games can be quite difficult, and pushing one’s limits is an excellent way to improve persistence, especially when accompanied by the great sense of satisfaction one gets upon successful completion of a very thorny problem (e.g., Eisenberg & Leonard, 1980; Eisenberg, 1992). Note, however, that some students may not feel engaged or comfortable with games, or cannot access them. For these students alternative approaches should be available.

As can be seen, traditional tests are not fully satisfying various validity and learning requirements. In the next section we describe how digital games can be effectively used in education—as assessment vehicles and to support learning.
Digital Games, Assessment, and Learning

Digital games are very popular. For instance, revenues for the digital game industry reached 7.2 billion in 2007 (Fullerton, 2008), and overall, 72% of the population in the U.S. plays digital games (Entertainment Software Association, 2011). The amount of time spent playing games also continues to increase (Escobar-Chaves & Anderson, 2008). Besides being a popular activity, playing digital games has been shown to be positively related to a variety of cognitive skills (e.g., visual-spatial abilities, Green & Bavelier, 2007; attention, Shaw, Grayson, & Lewis, 2005), personality traits (e.g., Openness, Chory & Goodboy, 2011; Ventura, Shute, & Kim, 2012; Witt, Massman, & Jackson, 2011), academic performance (e.g., Skoric, Teo, & Neo, 2009; Ventura, Shute, & Kim, 2012), and civic engagement (Ferguson & Garza, 2011). Digital games can also motivate students to learn valuable academic content and skills, within and outside of the game (e.g., Barab et al., 2010; Coller & Scott, 2009; DeRouin-Jessen, 2008). Finally, studies have shown that playing digital games can promote prosocial and civic behavior (e.g., Ferguson & Garza, 2011).

As mentioned earlier, learning in games has historically been assessed indirectly and/or in a post hoc manner (see Shute & Ke, in press). What is needed instead is real-time assessment and support of learning based on the dynamic needs of players. We need to be able to experimentally determine the degree to which games can support learning, and how and why they achieve this objective. Research examining digital games and learning is usually conducted using pretest-game-posttest designs, where the pre- and posttests measure content knowledge. Such traditional assessments don’t capture and analyze the dynamic and complex performances that inform 21st century competencies. How can we both measure and enhance learning in real time? Performance-based assessments with automated scoring are needed. The main assumptions
underlying this new approach are that: (a) learning by doing (required in game play) improves learning processes and outcomes, (b) different types of learning and learner attributes may be verified and measured during game play, (c) strengths and weaknesses of the learner may be capitalized on and bolstered, respectively to improve learning, and (d) feedback can be used to further support student learning.

**Evidence of Learning from Games**

Below are three examples of learning from digital games that represent commercial as well as educational games. Preliminary evidence suggests that students can learn deeply from such games, and acquire important 21st-century competencies.

**Deep learning in Civilization.** Our first example illustrates how a commercial digital game can be used to support deep learning of history. Kurt Squire, at the University of Wisconsin, used a strategy game called Civilization in a high school world history class (Squire, 2004). The goal of this game is to build, advance, and protect a civilization. This game starts with kids picking a civilization that they want to build (e.g., ancient Mesopotamia). Kids make many decisions about how to build and grow their civilization. Sometimes their decisions can be as simple as deciding where to put a new bridge, but they can be as complex as deciding whether to start a nuclear war. To make successful decisions, a player needs to consider important elements of human history, including economy, geography, culture, technology advancement, and war. A screen capture of the game Civilization is shown in Figure 1.
So what do kids learn from playing this game? Squire reported that players mastered many historical facts (e.g., where Rome was located), but more importantly, at the end of the game, they took away a deep understanding about the intricate relationships involving geographical, historical, and economic systems within and across civilizations.

**Gamestar Mechanic and systems thinking.** Our next example illustrates how digital games can be used to support systems thinking skill. Systems thinking skill refers to a particular way of looking at the world which involves seeing the “big picture” and the underlying interrelationships among the constituent elements rather than just as isolated bits. Gamestar Mechanic is an online game that is intended to teach kids basic game design skills and also allows them to actually build their own games for themselves, friends, and family to play. To design a functioning and challenging game in Gamestar Mechanic, players need to think deeply about various game elements, parameters, and their interrelationships. If they think too simply,
and just change a few elements of the game without considering the whole system, the game will not work. For example, consider a player who included too many enemies in her game (each one with full strength). The consequence of this decision would be that other players would not be able to beat the game, so it would not be any fun. With a little reflection, she would realize the impact that the number/strength of enemies feature of the game would have on other elements of the game, and revise accordingly. A screen capture showing some of the settings a player can manipulate in Gamestar Mechanic is shown in Figure 2.

![Screen capture of settings in Gamestar Mechanic](image)

*Figure 2: Screen capture of settings in Gamestar Mechanic (Zimmerman, 2010).*

Torres (2009) recently reported on his research using Gamestar Mechanic. He found that kids who played the game did, in fact, develop systems thinking skills along with other important skills such as innovative design.

**Taiga Park and science content learning.** Our last example illustrates how kids learn science content and inquiry skills within an online game called Quest Atlantis: Taiga Park. Taiga Park is an immersive digital game developed by Sasha Barab and his colleagues at Indiana University (Barab et al., 2007; Barab, Gresalfi, & Ingram-Goble, 2010). Taiga Park is a beautiful
national park where many groups co-exist, such as the fly-fishing company, the Mulu farmers, the lumber company, and park visitors. In this game, Ranger Bartle calls on the player to investigate why the fish are dying in the Taiga River. To solve this problem, players are engaged in scientific inquiry activities. They interview virtual characters to gather information, and collect water samples at several locations along the river to measure water quality. Based on the collected information, players make a hypothesis and suggest a solution to the park ranger.

To move successfully through the game, players need to understand how certain science concepts are related to each other (e.g., sediment in the water from the loggers’ activities causes an increase to the water temperature, which decreases the amount of dissolved oxygen in the water, which causes the fish to die). Also, players need to think systemically about how different social, ecological, and economical interests are intertwined in this park. In a controlled experiment, Barab et al. (2010) found that middle school students learning with Taiga Park scored significantly higher on the posttest (assessing knowledge of core concepts such as erosion and eutrophication) compared to the classroom condition. The same teacher taught both treatment and control conditions. The Taiga Park group also scored significantly higher than the control condition on a delayed posttest, thus demonstrating retention of the content relating to water quality. See Figure 3 for a screen capture from Taiga Park.
As these examples show, digital games appear to support learning. But how can we more accurately measure learning, especially as it happens (rather than after the fact), and beyond just content knowledge?

Assessment in Games

In a typical digital game, as players interact with the environment, the values of different game-specific variables change. For instance, getting injured in a battle reduces health and finding a treasure or another object increases your inventory of goods. In addition, solving major problems in games permits players to gain rank or “level up.” One could argue that these are all “assessments” in games—of health, personal goods, and rank. But now consider monitoring educationally-relevant variables at different levels of granularity in games. In addition to checking health status, players could check their current levels of systems thinking skill, creativity, and teamwork, where each of these competencies is further broken down into constituent knowledge and skill elements (e.g., teamwork may be broken down into cooperating,
negotiating, and influencing/leadership skills). If the estimated values of those competencies got too low, the player would likely feel compelled to take action to boost them.

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

ECD is a conceptual framework that can be used to develop assessment models, which in turn support the design of valid assessments. The goal is to help assessment designers coherently align (a) the claims that they want to make about learners, and (b) the things that learners say or do in relation to the contexts and tasks of interest (e.g., Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003, and for a simple overview, see ECD for Dummies). There are three main theoretical models in the ECD framework: competency, evidence, and task models.

**Competency model:** *What collection of knowledge, skills, and other attributes should be assessed?* Although ECD can work with simple one-dimensional competency models, its strength comes from treating competency as multidimensional. Variables in the competency model (CM) describe the set of knowledge and skills on which inferences are based (see Almond & Mislevy, 1999). The term student model is used to denote an instantiated version of the CM – like a profile or report card, only at a more refined grain size. Values in the student model express the assessor’s current belief about the level on each variable within the CM, for a particular student.
**Evidence model:** *What behaviors or performances should reveal those competencies?*

An evidence model expresses how the student’s interactions with, and responses to a given problem constitute evidence about competency model variables. The evidence model (EM) attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies; and (b) What’s the statistical connection between those behaviors and the CM variable(s)?

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

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

The second reason that ECD is believed to work well with digital games is because the ECD framework is based on the assumption that assessment is, at its core, an evidentiary
argument. Its strength resides in the development of performance-based assessments where what is being assessed is latent or not apparent (Rupp, Gushta, Mislevy, & Shaffer, 2010). In many cases, it is not clear what people learn in digital games. However in ECD, assessment begins by figuring out just what we want to assess (i.e., the claims we want to make about learners), and clarifying the intended goals, processes, and outcomes of learning.

Accurate information about the student can be used to support learning. That is, it can serve as the basis for (a) delivering timely and targeted feedback, as well as (b) presenting a new task or quest that is right at the cusp of the student’s skill level, in line with flow theory (e.g., Csikszentmihalyi, 1990) and Vygotsky's zone of proximal development (Vygotsky, 1978).

As discussed so far, there are good reasons for using games as assessment vehicles to support learning. However, Zapata-Rivera and Bauer (2011) discuss some of the challenges relating to the implementation of assessment in games. These challenges include the following:

- **Introduction of construct irrelevant content and skills.** When designing interactive gaming activities, it is easy to introduce content and interactions that impose requirements on knowledge, skill, or other attributes (KSA) that are not part of the construct (i.e., the KSAs that we are not trying to measure). That is, authenticity added by the context of a game may also impose demands on irrelevant KSAs (Messick, 1994). Designers need to explore the implications for the type of information that will be gathered and used as evidence of students’ performance on the KSAs that are part of the construct.

- **Interaction issues.** The nature of interaction in games may be at odds with how people are expected to perform on an assessment task. Making sense of issues such as exploring behavior, pacing, and trying to game the system is challenging and has a direct link to the
quality of evidence that is collected about student behavior. The environment can lend itself to interactions that may not be logical or expected. Capturing the types of behaviors that will be used as evidence and limiting other types of behaviors (e.g., “gaming” the system or repeatedly exploring visual or sound effects) without making the game dull or repetitive is a challenging activity.

- **Demands on working memory.** Related to both the issues of construct irrelevant variance (i.e., when the test contains excess variance that is irrelevant to the interpreted construct; Messick, 1989) and interaction is the issue of demands that game-like assessments place upon students’ working memory. By designing assessments with higher levels of interactivity and engagement, it’s easy to increase cognitive processing demands in a way that reduces the quality of the measurement of the assessment.

- **Accessibility issues.** Games that make use of rich, immersive graphical environments can impose great visual, motor, auditory, and other demands on the player to just be able to interact in the environment (e.g., sophisticated navigation controls). Moreover, creating environments that do not make use of some of these technological advances (e.g., a 3D immersive environment) may negatively affect student engagement, especially for students who are used to interacting with these types of games. Parallel environments that do not impose the same visual, motor and auditory demands without changing the construct need to be developed for particular groups of students (e.g., students with visual disabilities).

- **Tutorials and familiarization.** Although the majority of students have played some sort of video game in their lives, students will need support to understand how to navigate and interact with the graphical environment. Lack of familiarity with navigation controls may
negatively influence student performance and student motivation (e.g., Lim, Nonis, & Hedberg, 2006). The use of tutorials and demos can support this familiarization process. The tutorial can also be used as an engagement element (e.g., Armstrong & Georgas, 2006).

• *Type and amount of feedback.* Feedback is a key component of assessments that informs instruction. Research shows that interactive computer applications that provide immediate, task-level feedback to students can positively contribute to student learning (e.g., Hattie & Timperley, 2007; Shute, 2008; Shute, Hansen, & Almond, 2008). Shute (2008) reviews research on formative feedback and identifies the characteristics of effective formative feedback for specific situations (e.g., multidimensional, non-evaluative, supportive, timely, specific, and credible). Immediate feedback that results from a direct manipulation of objects in the game can provide useful information to guide exploration or refine interaction strategies. The availability of ongoing feedback may influence motivation and the quality of the evidence produced by the system. Measurement models need to take into account the type of feedback that has been provided to students when interpreting the data gathered during their interaction with the assessment system (e.g., one way of measuring the effects of feedback is by varying the difficulty of the activity based on the feedback provided).

• *Re-playing, number of attempts, and revisions.* As in the case of feedback, measurement models need to be able to handle any number of attempts and revisions of solutions during game play. This could be done by comparing the outcomes of consecutive actions/events or by interpreting a subset of actions. Based on the type of assessment,
operational constraints (e.g., time) may impose a limit on the number of attempts allowed before moving to the next scenario.

- **Handling dependencies among actions.** Dependencies among actions/events can be complex to model and interpret. Assumptions of conditional independence required by some measurement models may not hold in complex interactive scenarios. Designing scenarios carefully can help reduce the complexity of measurement models. Using data mining techniques to support evidence identification can also help with this issue.

In addition to these challenges, in order to make scalable assessments in games, we need to take into account operational constraints and support the need for assessment information by different educational stakeholders including students, teachers, parents and administrators.

Stealth assessment addresses some of these challenges. The next section describes stealth assessment and provides a sample application in the area of Newtonian physics.

Stealth Assessment

Given the goal of using well-designed games to support learning in school settings and elsewhere, we need to ensure that the assessments are valid, reliable, and also unobtrusive (to keep engagement intact). One way to meet these requirements is to use “stealth assessment” (Shute, 2011; Shute et al., 2009). Stealth assessment refers to ECD-based assessments that are woven directly and invisibly into the fabric of the learning or gaming environment. During game play, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess (e.g., scientific inquiry skills, creativity). Evidence needed to assess the skills is thus provided by the players’ interactions with the game itself (i.e., the processes of play), which can be contrasted with a typically singular outcome of an activity—the norm in educational environments.
Making use of this stream of gameplay evidence to assess students’ knowledge, skills, and understanding (as well as beliefs, feelings, and other learner states and traits) presents problems for traditional measurement models used in assessment. First, in traditional tests the answer to each question is seen as an independent data point. In contrast, the individual actions within a sequence of interactions in a game are often highly dependent on one another. For example, what one does in a particular game at one point in time affects subsequent actions later on. Second, in traditional tests, questions are often designed to measure particular, individual pieces of knowledge or skill. Answering the question correctly is evidence that one may know a certain fact: one question – one fact. But by analyzing a sequence of actions within a quest (where each response or action provides incremental evidence about the current mastery of a specific fact, concept, or skill), stealth assessments within game environments can infer what learners know and do not know at any point in time. Now, because we typically want to assess a whole cluster of skills and abilities from evidence coming from learners’ interactions within a game, methods for analyzing the sequence of behaviors to infer these abilities are not as obvious. As suggested above, evidence-based stealth assessments can address these problems.

**Stealth Assessment in Crayon Physics Deluxe**

We have designed a number of stealth assessment *mockups* for measuring competencies within different games, such as systems thinking skills in *Taiga Park* (Shute, Masduki, & Donmez, 2010), creative problem solving in *Oblivion* (Shute et al., 2009), and causal reasoning in the *World of Goo* (Shute & Kim, 2011). What needs to be done now is to actually build stealth assessments directly within a digital game, as part of gameplay. In a current research project, we are doing just that. Before describing the game and stealth assessments, we first describe the research project.
**Research Project.** Less than a year ago, we (first and second authors) received funding from the Bill and Melinda Gates Foundation to design, develop, and evaluate three evidence-based assessments embedded in a digital game. The three focal competencies are: creativity, conscientiousness, and conceptual physics understanding in the game Crayon Physics Deluxe (described in the next section). Data will be dynamically collected in CPD from players’ interactions in the game to inform our three focal competencies.

We are currently in the first year of the two-year project which involves creating three stealth assessments. We are using problems that already exist in CPD, as well as creating new ones with the game’s level editor, to suit our experimental needs (e.g., creating certain aspects of a problem, such as difficulty level and physics principles needed in the solution). We have begun pilot testing the problems to determine if they’re suitable for our population and for our methodological requirements (e.g., adequate variability). Pilot work is being conducted with a sample of middle school students at the Florida State University School (FSUS).

In our second year, we will conduct two studies to evaluate the validity of the stealth assessments, examine learning from the game, and test the scalability of the stealth assessments to other games. Specifically, in the first study ($n = 120$), we will evaluate the validity of our three stealth assessments in CPD. Students will (a) complete a pretest battery of traditional tests on our three focal competencies, (b) interact with a pool of CPD problems (20-30 carefully selected problems) over four 1-hour sessions in the computer lab at their school, and (c) complete a posttest on conceptual physics understanding. Students’ competency levels will be estimated from their gameplay in CPD, and the competency estimates will be correlated with scores from the traditional tests. The results of the study will inform us as to the validity of the stealth assessments for the three focal competencies (creativity, conscientiousness, and conceptual
physics) and provide us with preliminary evidence for conceptual physics learning in CPD.

In our second study during year 2, we will employ one of our stealth assessment models (i.e., persistence, which is a main facet of conscientiousness) in a different digital game (e.g., World of Goo). A subset of the students from the first study will be used in the second study ($n = 80$) to evaluate how persistence can be assessed in this second game. Students will interact with the second game over two 1-hour sessions in the computer lab at their school. At the end of the sessions, the competency estimates for persistence will be compared to the Study 1 estimates from CPD and to scores from the traditional tests to evaluate the validity of the assessments, and the scalability of the persistence models (i.e., developed for one game and reused within another one).

We now turn our attention to the game we are using as the vehicle for our stealth assessment project.

**Crayon Physics Deluxe (CPD).** CPD is a computer game that emphasizes two-dimensional physics simulations, including gravity, mass, kinetic energy, and transfer of momentum. The objective of each problem in CPD is to guide a red ball from a predetermined starting point to a star (or stars). Everything obeys the basic rules of physics relating to gravity and Newton’s three laws of motion. The player can nudge the ball to the left and right (if the surface is flat) but the primary way to move the ball is by drawing physical objects on the screen that “come to life” once the object is drawn. For example, in the “golf problem” (see Figure 4), the player must draw a golf club on a pin (i.e., little circle inside the cloud) to make it swing down to hit the ball. In the depicted solution, the player also drew a ramp to prevent the ball from falling down a pit.
The speed of (and importantly, the impulse delivered by) the swinging golf club is dependent on the size/mass distribution of the club and the angle from which it was dropped to swing. The ball will then fly at a certain speed, length, and trajectory. If drawn properly, the ball will hit the star.

The various problems in CPD require the player to create and use catapults, pulleys, pendulums, levers, and so forth to move the ball. Another screen capture is shown in Figure 5 with the ball on the left, the star on the right, and some gears/chains obstructing the way. All solutions are drawn with crayons using the mouse. In a number of cases the ball must go over a pit. If the ball falls into the pit, the player must start the problem over.
Players can replay a problem as often as they like—even after successfully solving it. One motivation to replay a problem is to find even more elegant and creative solutions than were generated before. It is not uncommon for a player to revisit/replay particularly challenging problems multiple times, always striving for a better, more elegant solution.

**Agents of force and motion.** As noted before, CPD requires players to create and use the following devices (or what we have been calling “agents of force and motion”) to help the ball reach the star:

1. **Ramp:** A ramp can be employed to change the direction of the motion of the ball (or another object). In some cases, other tools (like a pendulum or nudge), are needed to get the ball to start moving.

2. **Lever:** A seesaw or lever involves net torque. A lever rotates around a fixed point usually called a fulcrum or pivot point. An object residing on a lever gains potential energy as it is raised.
3. **Pendulum**: A swinging pendulum directs an impulse tangent to its direction of motion. The idealized pendulum is a specialized case of the physical pendulum for which the mass distribution helps determine the frequency. One can draw a physical pendulum in CPD, and the motion will be determined by the mass distribution.

4. **Springboard**: A springboard (or diving board) stores elastic potential energy provided by a falling weight. Elastic potential energy becomes kinetic as the weight is released.

5. **Pulley**: A pulley shows how a rope can change direction of force. One draws a rope going over a pulley. In CPD, ropes are effectively mass-less so that the force applied to one end of a taut rope is transmitted undiminished to the object on the other end (see **Rope**).

6. **Pin**: A pin allows the position of one body to be fixed in space. Like a nail, it supplies a force large enough to resist motion of the point it is attached to. Two pins hold a body immobile against a background.

7. **Rope**: Ropes generally transmit tension between objects. If a rope is draped over a pulley with masses attached at both ends and the masses are equal, their weights are equal and the net force on each will be the difference between the tension pulling up on the mass and the force of gravity pulling down. Ropes can also act like trampolines, generating forces on objects by stretching the rope and then removing the force (by deleting objects) to produce upward momentum on the ball.

8. **Nudge**: An arrow in CPD allows the user to poke/nudge an object into motion.

The next section introduces the three competency models and their associated indicators (i.e., evidence) in the CPD project. For each of the three CMs, we: (a) review the relevant literature, then (b) present a coherent graphical model of the variables. In the graphical models,
unobservable/theoretical variables are on the left and the specific observable/measurable indicators are on the right (i.e., what a person does in the environment to inform the latent variables).

Conscientiousness Review and CM

Over the past 20 years or so, conscientiousness has emerged as one of the most important competencies in academia (e.g., Proprat, 2009) as well as in the workforce (e.g., Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007; Schmidt & Hunter, 1998). Conscientiousness (C) is a multi-faceted competency that commonly includes tendencies related to being attentive, hard-working, careful, detail-minded, reliable, organized, productive, and persistent (Noftle & Robins, 2007; Roberts, Chernyshenko, Stark, & Goldberg, 2005). It is also important to note that C is not highly related to mathematics skill or verbal reasoning (Trapmann, Hell, Hirn, & Schuler, 2007), measures typically used to assess general cognitive ability. Thus C is considered to be noncognitive (i.e., a person’s level of C is relatively independent from cognitive measures such as standardized tests like ACT or SAT).

C’s independence from intelligence means it can affect students with high or low levels of cognitive ability. For example, a person who has high cognitive ability but low C may end up performing about the same in school as a person who is low on cognitive ability and high on C. Thus C can be seen as an independent ability that can help or hinder performance in school. It is unclear why certain people have higher or lower levels of C. C does not appear to be related to SES (Roberts et al., 2007) but has been shown to increase over one’s lifetime (Roberts, Walton, & Viechtbauer; 2006). The next section reviews the empirical evidence regarding the validity of C.

Validity of Conscientiousness. A number of studies and meta-analyses have shown the
importance of self-report measures of C in predicting a variety of important outcomes while controlling for cognitive ability. C has consistently been found to predict academic achievement from preschool (Abe, 2005) to high school (Noftle & Robins, 2007, Proporat, 2009), to the postsecondary level (O’Conner & Paunonen, 2007; Trapmann, Hell, Hirn, & Schuler, 2007) and adulthood (e.g., De Fruyt & Mervielde, 1996; Shiner, Masten, & Roberts, 2003). Meta-analyses have linked C with grades between $r = .21$ and $.27$, and the relationship is independent of intelligence (e.g., Noftle & Robins, 2007, Proporat, 2009; Robbins Lauver, Le, Davis, Langley, & Carlstrom, 2004). Existing research suggests that the process-oriented aspects of C (e.g., organizing and planning) show the weakest relationships with school achievement while the aspects representing goal completion, achievement, and productivity show the strongest relationships (e.g., industriousness, achievement striving, achievement motivation; Noftle & Robins, 2007; Robbins et al., 2004). The next section describes the research on the structural facets of C.

**Structural Models of Conscientiousness.** MacCann, Duckworth, and Roberts (2009) reviewed studies that examined the underlying structure of C (Peabody & De Raad, 2002; Perugini & Gallucci, 1997; Roberts, Bogg, Walton, Chernyshenko, & Stark, 2004; Roberts, Chernyshenko, Stark, & Goldberg, 2005; Saucier & Ostendorf, 1999). These studies are summarized in Table 1.
Table 1. Underlying Structure of Conscientiousness from six studies (MacCann, Duckworth, & Roberts, 2009)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Unit Analyzed</th>
<th>Analysis</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
<th>Factor 8</th>
<th>Factor 9</th>
<th>Factor 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saucier &amp; Ostendorf (1999)</td>
<td>Adjectives ratings</td>
<td>EFA</td>
<td>Superficiality</td>
<td>Industriousness</td>
<td>Responsibility</td>
<td>Impulse Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roberts et al. (2004)</td>
<td>Adjectives ratings</td>
<td>EFA</td>
<td>Recklessness</td>
<td>Industriousness</td>
<td></td>
<td>Impulse Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roberts et al. (2005)</td>
<td>Self-report ratings</td>
<td>EFA</td>
<td></td>
<td></td>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MacCann, Duckworth, &amp; Roberts, 2009</td>
<td>Self-report ratings</td>
<td>EFA/CFA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The unit of analysis is important to distinguish since some studies conducted factor analysis on adjective ratings while others used factor analysis on self-report ratings. Three factors were common to all five studies (orderliness, industriousness, and responsibility/reliability), a control factor emerged in four of the five studies, and decisiveness and conventionality factors emerged in two studies.

Regarding the MacCann et al. (2009) study, confirmatory factor analysis uncovered eight facets: industriousness, perfectionism, tidiness, proactivity, control, cautiousness, task planning, and perseverance. All facets related meaningfully to broad C, while perseverance also overlapped with neuroticism. The facets of industriousness and proactivity showed higher prediction of student absences and attainment of academic honors compared with the other
MacCann et al. (2009) discuss why facets are important for developing intervention programs targeting personality development. Specifically, facets are needed for diagnoses of specific deficiencies. Based on Table 1, we developed a competency model that is displayed in Figure 6.

As can be seen, we refined the C model to include only four facets: *persistence*, *perfectionism*, *organization*, and *carefulness*. Our persistence facet combines the industriousness and perseverance facets since they both imply the notion of “working hard despite failure.” We kept perfectionism as a facet, created the organization facet, and then broke it down into two
main sub-facets: time management and resource management. Time management can be broken down further to: proactivity (i.e., getting work done in a timely fashion) and task planning (i.e., the ability to plan ahead to meet deadlines). We created the carefulness facet which can be broken down into caution (i.e., being careful not to make mistakes) and control (i.e., tendency to not act impulsively). Finally, the figure includes CPD indicators that can linked to each of the facets. Based on the constraints of CPD, only persistence and perfectionism will be modeled.

**Task Modeling for Persistence.** Assessing persistence is primarily based on seeing how long players spend or persist on problems that they do not readily solve. The challenge in this assessment design is that we are never really sure what problem a student may or may not be able to solve. To address this issue, we created difficulty rubrics for problems to systematically manipulate problem difficulty. This allows us to incrementally increase the difficulty of problems to ensure that students will eventually get to problems they will have trouble solving. Difficulty indices include the following:

1. *Relative location of ball to star.* If the star is positioned above the ball in a problem, this forces the player to use a lever, pulley, springboard, or pendulum to solve the problem (0-1 point).

2. *Obstacles.* This refers to the pathway between the ball and star. If the pathway is obstructed, this requires the player to project the ball in a very specific trajectory to obtain the star (0-2 points).

3. *Distinct agents of force/motion* (see previous section on Agents of Force and Motion). A CPD problem may require one or two agents to get the ball to the star (0-1 point).
4. **Novelty.** This addresses whether a problem is novel relative to other problems played. Problem solution is not easily determined from experience with other problems (0-2 points).

Each problem is evaluated under all of the rubrics to yield a total difficulty score (i.e., ranging from 0-6). Consider the problem “cave story” in Figure 7. This problem gets a difficulty score of 5 as the star is above the ball (1), there’s one obstacle which is a narrow pathway (1), two agents are typically needed to solve it (1) and there’s no other problem like that in the game (2). Thus “cave story” would be a good problem to assess persistence as it will likely be unsolvable by many students.

![Figure 7: Cave story in CPD](image)

**Creativity review and CM**

Creativity is generally defined as the ability to produce solutions or ideas that are both novel and effective (Lubart, 1994). Kaufman and Sternberg (2007) have noted that most definitions of creativity (or creative problem solving) consist of three components: novelty, quality, and relevance. That is, creative solutions are novel, of high quality, and appropriate to the given task, or some variant of the task.
Various psychometric approaches exist to help understand and model creativity. Sternberg and Lubart (1992) explained that the different approaches to creativity can be viewed as the continuum between “less” contextualized approaches that focus on personal characteristics, and “more” contextualized approaches that include social-cultural variables that influence individuals’ creativity. Typical “less” contextualized psychometric approaches explain creativity as a multifaceted construct that includes intelligence, knowledge, thinking styles, and personality traits. McCrae (1987) stresses that the ability to think creatively in conjunction with an inclination to do so (i.e., disposition) leads to creative productions. Many other creativity researchers share similar views that creativity is a multifaceted construct that involves a convergence of multiple variables (e.g., Amabile, 1983).

**Validity of Creativity.** Creativity has probably been one of the most elusive of all scientific constructs in psychology. For the last 30 years, the field (of creativity research) has received criticism challenging its validity (Weisberg, 1993). Nevertheless, many published reviews in creativity and innovation research show that interest in creativity is strong and is increasingly viewed as a key component relevant to academic success (e.g., Gronhaug & Kaufman, 1988; Kaufman, 2003; Sternberg, 1988, 1999, 2006; Runco, 1997, 2002; Runco & Pritzker, 1999).

Divergent thinking tests are among the most popular techniques for measuring creativity in educational settings (Hunsaker & Callahan, 1995; Runco, 1992). These tests, also referred to as measures of ideational fluency, generally require students to provide as many responses as possible to prompts such as, "List things that make noise" or "List things that have wheels." Among the most popular of the creativity tests are (a) the Torrance Tests of Creative Thinking (Torrance, 1974) and (b) the Wallach and Kogan (1965) tests. Responses are usually scored for
originality by expert raters and fluency (number of responses). These test have been shown to moderately predict important criteria such as school success (e.g., Okuda, Runco, & Berger, 1991; Runco & Pritzker, 1999; Sternberg, 2006).

**Structural Models of Creativity.** Guilford (1956) conceptualized creativity as involving four facets of divergent thinking—fluency (the ability to rapidly produce a large number of ideas), flexibility (the ability to produce ideas from various categories or classes), originality (the ability to produce ideas that are unique, novel, and uncommon), and elaboration (the ability to develop the details of an idea and carry it out). Flexibility has been recognized as an essential cognitive skill for creativity (Amabile, 1983) and is defined as the ability to generate a varied pool of ideas by switching among categories and using remote associations (Nijstad, De Dreu, Rietzschel, & Baas, 2010; Runco, 1986).

Runco (1986) discusses the significance of flexibility for creativity assessment: (a) it distinguishes gifted children from non-gifted better than fluency and originality, and (b) it contributes to the predictive validity of divergent thinking tests with real-world criteria. People with a higher level of flexibility avoid using fixed problem solving strategies, break perceptual sets, and make new connections among distant ideas. Even though the cognitive skills that are required for ideation (i.e., divergent thinking) are often considered as being synonymous with creativity, many caution that divergent thinking explains only one aspect of creativity, not the whole (e.g., Runco, 2008). We agree.

Openness to experience, one of the dimensions of the Big-Five factors, refers to a dispositional attribute that is characterized by an awareness of personal feelings and beliefs, receptivity to novel ideas, liberal values, intellectual curiosity, and fantasy (Berzonsky & Sullivan, 1992). Therefore, individuals with higher degrees of openness to experience are
described as imaginative, sensitive to aesthetics, curious, independent thinkers, and/or amenable to new ideas, experiences, and unconventional views (Costa & McCrae, 1992; McCrae, 1996; McCrae & Costa, 1997). Torrance (1974) explains that a creative individual tends to resist premature closure by keeping herself open-minded and considering a variety of information sources. A long line of research has supported the strong association of openness to experience with creativity or some aspects of creativity (Costa & McCrae, 1992; Dollinger, Urban, & James, 2004; Feist, 1999; George and Zhou, 2001; McCrae, 1987, 1996; McCrae & Costa, 1997). For example, McCrae (1987) reported a significant association ($r = .4$) between divergent thinking and openness to experience.

Willingness to take risks (i.e., risk propensity) can be defined as the extent to which an individual takes an action knowing there is uncertainty related to the potential pay-off of the action (Dewett, 2007). Risk-taking is associated with openness to change and new ideas (Madjar, Greenberg, & Chen, 2011). Willingness to take risks (and knowing the possibility of failing) has been recognized as an essential trait of eminent scientists and artists throughout history. Sternberg and Lubart (1992) describe creative individuals as those who “buy low and sell high.” They further argue that willingness to take risks is prerequisite for growth and creativity because one needs to go beyond what is commonly accepted and learn from various failings. Several studies have reported a positive association between willingness to take risks and creativity (Glover, 1977; Glover & Sautter, 1977). For example, Glover and Sautter (1977) reported that willingness to take risks was significantly correlated with flexibility and originality. Willingness to take risks also has been studied in the context of organizational innovations for many years (e.g., Dewett, 2004, 2007; Kogan & Wallach, 1964; MacCrimmon & Wehrung, 1990).
example, Madjar, Greenberg, and Chen (2011) reported that willingness to take risks is a significant contributor to individuals’ radical creativity and innovation.

Based on the literature, we have developed a CM of creativity as shown in Figure 8. As can be seen, the model broadly splits creativity into Cognitive and Dispositional facets. The Cognitive facet primarily refers to the ability to be creative in problem solving tasks and is knowledge dependent. The dispositional facet refers to creative dispositions (e.g., openness) that are not necessarily related to cognitive ability or domain. Within the cognitive facet there are four sub-facets: fluency, flexibility, originality, and elaboration. Based on the constraints of CPD we will only focus on the cognitive facet of creativity.

![CM of Creativity with indicators from CPD](image)

*Figure 8: CM of Creativity with indicators from CPD*
**Task Modeling for Originality.** Originality is assessed by seeing how unique a player’s solution is relative to other students’ solutions on a particular problem. This can be determined by seeing what agent(s) of force and motion a student used in a solution, or what trajectory the ball traveled in a solution. To ensure that we do, in fact, see variation in player solutions, we are creating problems (using the game’s level editor) that can be solved with many different agents. Consider the problem that we call *four agents* (see Figure 9).

![Four agents in CPD](image)

*Figure 9: Four agents in CPD*

This problem can be solved with the lever, pulley, pendulum, and/or the springboard. What we expect to see is that players will solve the problem (and others like it) in multiple and varied ways. This will be encouraged via the instructions we will give to the students in the experiment. The instructions are:

*You will have 60 minutes to solve 10 problems. Your goal is to solve as many of the problems, in as many awesome ways as you can. The tools we taught you will come in handy for many problems. However, feel free to solve any problem in whatever way you like. You also have the freedom to jump around and solve*
problems in any order that you like. For example, if you get stuck on a problem, you can leave it and come back to it later.

Each problem can be solved with one or more of your new tools. Each time you solve a problem (collect a star) with a different tool, you’ll be given a flag. If you solve a problem with just one object (which is hard) you get a special flag. Your goal is to have as many flags as possible. Again, if you get stuck on a problem, you can always leave that problem and go to another problem. Just press the "escape" key and follow the directions to exit the problem. You can return to any problem as many times as you like. If you have any questions about how to play the game, please ask. And remember to view the short videos to remind you how a tool works in the game. Have fun!

When each one-hour “play” session is complete, we will be able to calculate the frequency of each agent of force and motion per problem. With this frequency information, we can see how common (or infrequent) a student’s agent was for a particular problem. Solutions that use infrequently-employed agents indicate originality. Next we discuss our model for conceptual physics understanding.

Conceptual Physics Review and CM

Physics engines are becoming pervasive in gaming environments, providing a sense of realism in a game (e.g., Havok engine). Within these gaming environments, players can experiment with principles of physics such as impulse, inertia, vector addition, elastic collision, gravity, velocity, acceleration, free-fall, mass, force, and projectile motion. The degree that players apply these principles correctly in the game can be evidence for conceptual understanding of physics. Specifically, players successfully drawing and applying the CPD agents of force and motion during problem solution will provide evidence related to knowing associated physics principles.

Over the past several decades it has become very clear that many students, who have gotten acceptable grades in one or more physics courses, actually have very limited understanding of the physics involved. Halloun (1996) traces this complaint back to Swann
(1950), and it has become a recurrent theme in the physics education research literature, even though many college and university physics faculty members prefer to “conduct business as usual” rather than entertain the possibility that serious changes in their own teaching might be needed.

The work of Mazur and others demonstrates that there are a number of routes to a passing grade that fail to develop an appreciation of physical principles and, more importantly, do not remove erroneous notions of how the world works from the students’ understanding (e.g., Crouch & Mazur, 2001; Hake, 1998; Halloun & Hestenes, 1985; McDermott, 1993). This has led to widespread adoption of the text Conceptual Physics by Paul Hewitt, currently in its 11th edition (2009), and the development of two instruments, the Force Concept Inventory (Hestenes et al., 1983) and the Mechanics Baseline Test (Hestenes & Wells, 1992) now widely used to compare student mastery of the concepts of mechanics between instructional approaches and courses. Recognition of the problem has also led to a renewed interest in the mechanisms by which physics students make the transition from naive or folk physics to Newtonian physics (di Sessa, 1982) and to the possibility of video game playing assisting in the process (White, 1994).

Based on foundational conceptual physics (e.g., Feynman, 1964; Feynman, Leighton, Sands, 1964; Hewitt, 2011), we interpret competency in conceptual physics to involve the following:

1. **Conceptual understanding of Newton’s three laws of motion.** Newton's three laws of motion provide a conceptual understanding of how objects interact in the environment. The first law tells us that an object in rest stays at rest in the absence of any unbalanced forces, and an object in motion stays in motion in a straight line with unchanging speed in the absence of any forces. The second law \( F = ma \) tells us how the motion of the particle (object) evolves when it experiences a nonzero net force. Here \( F \) is the net force applied (i.e., the vector sum of all the forces acting on the object), \( m \) is the mass of the object, and \( a \) is the object’s
acceleration. Thus, the net force applied to an object produces a proportional acceleration. That is, if an object is accelerating, then there is a net nonzero force on it. Any mass that is gained or lost by the system will cause a change in momentum that is not the result of an external force. In simple terms, it takes less force to produce the same acceleration of an object that has less mass compared to one with more mass. The third law states for every action there is an equal and opposite reaction. This is commonly described by hitting a tree with a baseball bat. The force exerted on the tree by the swinging bat is equal to the force exerted back on the bat by the tree.

2. **Conceptual understanding of potential and kinetic energy.** Potential energy exists when a force acts on an object to restore the object to its resting point (or “lower energy configuration”). For example, when a springboard (like in CPD) is bent downward, it exerts an upward force to return to its un-bent position. The action of bending the springboard down requires energy, and the work done by the springboard in returning it to its resting point is considered stored as potential energy. When the bent springboard is released, the stored energy will be converted into kinetic energy.

3. **Conceptual understanding of conservation of angular momentum or torque.** The angular momentum of a system of objects about any point of reference can be computed from the position and momentum of each of the objects. A useful mental image is that of a figure skater or gymnast. Figure skaters will begin an elegant spin with arms outstretched. Once they start spinning, they typically draw their hands inward so that they can spin more rapidly. The sum of the mass of each object making up the skater times the square of the (perpendicular) distance to the axis of rotation is called the skater’s moment of inertia. For a rotating object, the angular momentum is the product of the moment of inertia and the angular velocity. With negligible friction between skater and ice, decreasing the moment of inertia by moving the arms inward increases the rotational velocity. Similar considerations apply to a gymnast doing somersaults while dismounting, or a diver on the way down to the water. A torque with a short moment arm can counterbalance the torque exerted by a much smaller force with a larger moment arm and vice versa. Consider an ordinary lever. The force of support at the fulcrum is not directly given, but the relation between torque and angular acceleration can easily be exploited by measuring torques from the fulcrum.
Figure 10 shows the version of our competency model for conceptual physics as it pertains to CPD (see Appendix 1 for the full model of physics principles). As can be seen, the model includes: Newton’s three laws, potential and kinetic energy, and conservation of angular momentum or torque.

Newton’s three laws is a parent principle in the model since it is pervasive in almost all problems in CPD. Successful use of each agent is an indicator of Newton’s three laws. Additionally, there are micro-indicators that inform each agent and principle as well. Table 2 displays our current set of micro-indicators.
Table 2. Micro-indicators for CPD agents of force and motion

<table>
<thead>
<tr>
<th>Agents</th>
<th>Micro-Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramp</td>
<td>1. Number of bends (or tubes, i.e., tortuosity)</td>
</tr>
<tr>
<td></td>
<td>2. Angle of each bend</td>
</tr>
<tr>
<td></td>
<td>3. Length of ramp</td>
</tr>
<tr>
<td>Lever</td>
<td>1. Length of the lever</td>
</tr>
<tr>
<td></td>
<td>2. Position of fulcrum</td>
</tr>
<tr>
<td></td>
<td>3. Height through which object falls before hitting lever</td>
</tr>
<tr>
<td></td>
<td>4. Mass of object</td>
</tr>
<tr>
<td></td>
<td>5. Location of the dropped object on lever (distance from fulcrum)</td>
</tr>
<tr>
<td>Pendulum</td>
<td>1. Angle of pendulum relative to horizontal fulcrum (180 degrees is max)</td>
</tr>
<tr>
<td></td>
<td>2. Length between the axis point and the fulcrum (Moment of inertia)</td>
</tr>
<tr>
<td></td>
<td>3. Mass (important when the pendulum hits something)</td>
</tr>
<tr>
<td></td>
<td>4. Position of pin</td>
</tr>
<tr>
<td>Springboard</td>
<td>1. Length of springboard</td>
</tr>
<tr>
<td></td>
<td>2. Mass of the object to weight it down</td>
</tr>
<tr>
<td></td>
<td>3. Position of the ball at release</td>
</tr>
<tr>
<td></td>
<td>4. Delete object or let fall off springboard</td>
</tr>
<tr>
<td></td>
<td>5. Angle of springboard at release (90 degrees max)</td>
</tr>
<tr>
<td>Pulley</td>
<td>1. Length of rope</td>
</tr>
<tr>
<td></td>
<td>2. Mass of counterweight</td>
</tr>
<tr>
<td></td>
<td>3. Height of the platform of the counterweight</td>
</tr>
<tr>
<td></td>
<td>4. Removal of platform</td>
</tr>
</tbody>
</table>

**Task Modeling for Conceptual Physics.** As with the other competencies, existing CPD problems as well as the problems we have created require the player to use one or more agents of force and motion in the solution. Successful solutions thus inform one or more of the competencies that we hope to develop in the student. As an illustration, consider the problem called *ballistic pendulum*, shown in Figure 11.
Figure 11. Ballistic pendulum problem

This problem requires the student to create a pendulum shape and initial angle with the vertical part positioned so that the pendulum will fall down and “kick” the ball into a free-fall trajectory that ends up landing on the star. Successfully solving this problem, in line with the CM, suggests that the student has an intuitive grasp of the concepts of torque, linear and angular momentum. Table 2 shows the micro-indicators that make up the pendulum problems. Implementing all of the micro-indicators is needed to successfully get the ball to hit the target. Incidentally, the ballistic pendulum is an experiment often done in introductory physics courses in high school or college.

The springboard is a variant of the lever in which one flat board rests on another object that is pinned in place, but hangs over one edge. Figure 12 shows the problem called diving board. When a weight is dropped from a height (or affixed) onto the free end of the springboard, the edge acts as an instantaneous axis of rotation and the board experiences an angular acceleration. This can be used to launch objects up into space. This requires knowledge of
potential and kinetic energy, and conservation of angular momentum. Table 2 shows the micro-indicators that make up a springboard problem. Again, correctly applying the micro-indicators is needed to successfully get the object launched into the tunnel to reach the star.

Figure 12. Diving board problem

Relation of physics indicators to conscientiousness and creativity indicators.

The indicators of conceptual physics understanding differ from those of the other competencies in that they must be experienced and learned. Also, they are domain specific. A measure – such as the number of attempts to solve a problem – might indicate a high level of persistence, but may also be consistent with a lesser understanding of physics. In addition, really creative, single-object solutions in CPD may come about through insight into physical principles or more simply by extensive trial and error.

The way we plan to resolve these issues is to model all relationships, within and among the three CMs, with Evidence Models. We are using Bayesian networks to establish the
conditional probability relationships among the variables within each CM and some of the relationships, like described above, between CMs.

**Capturing performance data.** So how do we capture the performance data that comes from game play, and use it to inform our three competency models? We are currently embedding code within the game that uses relevant game-play data to automatically generate evidence indicators. The complexity of this code varies depending on the indicator being generated. For example, it is simple to generate quantitative indicators such as time spent on a problem (persistence) or number of problem re-starts (elaboration). The aforementioned indicators can be generated using a *game timer* variable and *counter* variable, respectively.

Calculations for some of the other indicators, such as the particular agents of force and motion used in a problem, will be more complex. These indicators require multiple queries to the physics engine used to model the game physics, and a heuristic analysis of gameplay events. For example, consider the pendulum agent. To test if a pendulum is striking the ball, our scheme: (a) queries the physics engine to determine if there are any objects contacting the ball, (b) checks if any of these objects are attached by a single pin (i.e., allowing the object to freely rotate), then (c) determines if there is any change to the ball’s trajectory resulting from the contact. If all three criteria are satisfied, our scheme generates a pendulum strike indicator.

To gauge the accuracy of our scheme during pilot testing, we will choose a random selection of game-play sessions and utilize the “replay” feature of the game to perform a manual (visual) analysis of the indicators exhibited in each session. We will then compare the indicators generated from the manual analysis with those automatically generated by our stealth assessment scheme. Modifications to the code for the automatically-generated indicators will be made to
align with human classifications, and eventually indicator classifications below 85% accuracy (relative to human evaluation) will not be implemented.

**Example of an Evidence Model for creativity.** In general, the functional relationships among the competency model and evidence model (i.e., indicators—obtained automatically via code in the game) can be presented as conditional probabilities by using a Bayesian network approach. To illustrate, in our current model for creativity, the marginal probability of each level of the competency variables is initially set to roughly 33%, which is “uninformative” (see Figure 13). In some cases, like the number of drawn objects, we’d expect elegant (i.e., single object) solutions to be less frequently occurring as they are difficult to achieve (which influences our prior estimates). The difficulty and discrimination values of the indicators are also initially set to intermediate values because we do not yet have empirical data to know how those indicators actually function in our assessment, even though we have established some “difficulty indices” based on problem characteristics. Once we collect students’ data, the probability distributions will change, specific to our population.

![Figure 13. CM and EM for creativity—prior probabilities](image)
To illustrate how the Bayesian network accumulates evidence and passes the information to the student model (i.e., the CM that is specific to a student), we provide an example of a student’s performance in the game (see Figure 14 below).

The probabilities in the model will be refined based on responses from many students (i.e., many hours of game playing data). Subsequently, instantiation of this student’s evidence is used to infer values for latent variables. So, after two hour-long sessions with the game, this student has generated, on average, three or more solutions per problem (see grey rectangle on lower left of figure). However, she only used, on average, two specific agents per solution attempt (i.e., pulley and ramp—for successful and unsuccessful solutions). She has also been judged to be at a “medium level” regarding the average number of objects created, for both solved and unsolved problems, relative to the population’s performance data. All of these fluency indicators (which are measurable) provide information about the number of solutions and objects that a student creates. In other words—the more solutions, the more “fluent,” in line with our definition of one of the main facets of creativity.
Once this evidence of her performance is incorporated into the student model, note that her level on the fluency node shifts toward the medium level (i.e., increasing from 33% to 51.1%). Moreover, the estimation about her overall creativity inches toward medium (i.e., increasing from 33% to 40.2%), while estimations for the other facets (i.e., flexibility, originality, and elaboration) do not change much. Furthermore, an operational version of this model may include a variety of indicators not included here, such as indicators generated from pattern analysis processes from trace data across students and tasks, and indicators that provide evidence for more than one latent variable.

**CPD study procedure.** One of the challenges of conducting an assessment study with a game is that it requires the player to be comfortable with the mechanics of the game. In order to speed up this familiarization process, we have developed introductory videos as tutorials to teach students about the various agents of force and motion. These tutorials will be given to students at the beginning of the session, and will dynamically illustrate how to draw each agent to solve a simple problem. Later, during gameplay, students have the option to watch any agent-drawing video at any time. Once the students have been trained with the agents, we will begin testing knowledge about conceptual physics by administering problems that test knowledge about each agent of force and motion. For example, in the *lever problem*, the student can only use a lever to get the ball to the star. After the students have completed two problems for each agent (10 problems) they will proceed to the problems that elicit indicators for creativity and persistence (10 problems). In these problems, the student will have the option to skip around the 10 problems and solve each problem in whatever way they like. Some of these problems will be difficult and may not be solved by all students.
Discussion and Future Research

Traditional assessments are often too simplified, abstract, and decontextualized to suit current education needs. We need new assessments that measure what students actually can do with the knowledge and skills obtained inside and outside of school (Shute, 2009). Our proposed solution to this problem is stealth assessment within well-designed digital games which we believe can provide meaningful assessment environments by providing students with problems that require the application of various competencies. The first and most important step of this effort will be the determination of the validity of our stealth assessments. If, in fact, the stealth assessments accurately estimate the focal competencies, then the next logical step is to examine scalability. That is, what are the costs and benefits of reusing ECD-based models in different games to assess the same kinds of competencies? These issues are being examined in our research project described herein.

If we find that this particular stealth assessment methodology yields valid and reliable information, we plan to make the process and models broadly available to the community so that the work will continue and grow. The research can expand in a number of directions. First, we (and/or others) can explore the development of stealth assessments for additional competencies that have been shown to play important roles in academic (or life) success (i.e., communication skills, empathy, critical thinking skills, and so on). Second, we can explore the development of stealth assessments relating to content that is more directly aligned with the Common Core Standards (e.g., mathematics modeling, probability, or reading comprehension). Third, we can push the bounds of our stealth assessments relative to implementing the models in additional digital games as well as other digital learning environments to determine the range of
environments that may employ the same competency and evidence models, for a scalable, cost effective and engaging solution to the assessment of complex competencies.

Stealth assessment also has the potential to be useful for diagnostic purposes due to the fine-grained analysis of student behavior in situated contexts. In addition, real-time information about player competency states can be useful to support learning through hints and feedback, as well as dynamic matching of game difficulty level to player ability (e.g., providing more challenging problems for those with high levels of various skills).

One final note should be made about those who may view stealth assessment as a way to trick students. The main purpose of stealth assessment is to provide formative rather than summative assessment information—to support learning. Since stealth assessments are so closely tied to the performance in the game, any assessment data collected can inform the game to adjust difficulty levels for scaffolding players. In CPD this can be accomplished by assessing knowledge of physics principles (though scoring micro-indicators), and presenting problems at varying difficulty levels to help players understand the principles.

**Acknowledgements**

We would like to sincerely thank the Bill and Melinda Gates Foundation for their funding for this project, particularly Emily Dalton-Smith and Robert Torres. We would also like to thank the other members of the grant team—Yoon Jeon Kim, Don Franceschetti, Russell Almond, Matt Small, and Lubin Wang—for their awesome support given to the project.
References


Dollinger, Urban, & James, 2004


George and Zhou, 2001


Hestenes et al., 1983


Kellaghan & Madaus, 1991


Madaus & O’Dwyer (1999)


McCrae, 1996

McCrae & Costa, 1997


NCLB, 2002


Squire, 2004


Vorderer & Bryant, 2006

Walberg, & Stariha, 1992


Appendix 1: Full Physics Competency Model

- **Conceptual Physics**
  - **Particle kinematics**
    - Relationships between position, displacement, velocity, and acceleration
    - Trajectory of body in free fall is a parabola
    - Trajectory of body in inverse square gravitational field is an ellipse
  - **Forces and torques**
    - Forces add as vectors; Net force is sum of forces
    - Normal force between solids in contact
    - Use gravitational force to increase velocity
    - Normal force between solids in contact
    - Use gravitational force to increase velocity
    - Impulse depends on speed and mass
    - Trajectory of body in free fall
    - Force proportional to strain of object
    - Tension transmitted undiminished through rope
    - Normal force between solids in contact
    - In the absence of external forces total linear momentum is conserved
    - Impulse delivered to object is change in its momentum
    - For path independent forces can define potential energy so that total energy equal to kinetic plus potential is conserved
    - Path dependent forces are dissipative
  - **Conserved quantities**
    - Total mass = sum of masses of components
    - Moment of inertia = sum of mass x moment arm squared
    - Angular impulse is change in angular momentum
    - Torque depends on force and distance from axis
    - Force proportional to strain of object
    - In the absence of external torques, total angular momentum is conserved
    - Angular impulse delivered to object is change in angular momentum
    - Torque depends on force and distance from axis
  - **Extended objects**
    - Strain defined by Hooke’s Law
    - Relationship between orientation, angular velocity, angular acceleration
    - Torque depends on force and distance from axis
Appendix 2: Technical description of the open source code in the appendix