Proposal Overview

This proposal describes our plans to design, develop, and evaluate evidence-based assessments embedded in a digital game. Embedding assessments within games provides a way to monitor and assess a player’s current level on valued competencies in real time. During gameplay, students produce rich sequences of actions while performing complex tasks, drawing on a variety of competencies. Evidence needed to assess the competencies is thus provided by the players' interactions with the game itself (i.e., the processes of play). In this proposed research, we focus on developing and evaluating assessments for three focal competencies: creativity, persistence, and conceptual physics in the game *Crayon Physics Deluxe*. *Crayon Physics Deluxe* (CPD) is a computer-based game that emphasizes two-dimensional physics simulations, including gravity, mass, kinetic energy, and transfer of momentum. Data will be collected in CPD from players' interactions in the game to inform our three focal competencies. A key feature of embedded (or stealth) assessments is evidence-centered assessment design (ECD), which requires a systematic analysis of the assessment argument, including the claims to be made about the learner(s) and the evidence that supports those claims (Mislevy, Steinberg, & Almond, 2003; Shute, 2011).

Year 1 will involve creating three stealth assessments that will be embedded in CPD. We will be using problems that already exist in CPD, as well as creating new ones with the game's editor, to define parameters in our task modeling effort (i.e., aspects of the problems, such as difficulty level and physics principles needed in the solution). We plan to pilot test the problems to determine if they're suitable for our population and for our methodological requirements (e.g., adequate variability). Pilot work will be conducted with 200 middle school students at the Florida State University School (FSUS; where we have permission by the school to conduct this research) and possibly the Florida Virtual School (FLVS). In Year 2, we plan to 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. Study 1 \((n = 120)\) will evaluate the validity of our three stealth assessments in CPD. Students will (a) complete a pretest battery of traditional tests on our focal competencies, (b) interact with 15 CPD problems over three 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, persistence, and conceptual physics) and provide us with preliminary evidence for conceptual physics learning in CPD. In Study 2, we will employ our ECD models in a different digital game (e.g., *World of Goo*, *Swarm*, or *Super Meat Boy*). A subset of the students in Study 1 will be used in Study 2 \((n = 80)\) 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 competency estimates from CPD and to scores from the traditional tests to evaluate the (a) validity of the assessments, and (b) scalability of the models (i.e., developed for one game and reused within another one).
Project Description

The world is changing rapidly. Education is not. Preparing our kids to succeed in the 21st Century requires fresh thinking on how to design new kinds of assessments that are valid and also support learning of 21st Century competencies. A century ago, a person who acquired basic reading, writing, and math skills was considered to be sufficiently literate. The goal back then was to prepare young people for production jobs because 90% of the students were not expected to seek or hold professional careers (Shute, 2007). But when faced with highly technical and complex problems in today’s world, the ability to solve ill-structured problems and think creatively, critically, collaboratively, and systemically is essential (e.g., Shute & Torres, in press; Walberg, & Stariha, 1992).

Except in rare instances, our current education system neither teaches nor assesses these new 21st Century 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 (Nofite & 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. They also fail to assess what students actually can do with the knowledge and skills acquired in, and often 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. 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 types of assessment methods are needed to properly assess valued competencies (Shute, 2009). Students need to be assessed in engaging, situated environments (like digital games) rather than by having to fill in bubbles on a prepared test form.

There are two main reasons for using well-designed games as vehicles for assessment in addition to their being engaging environments. First, 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, & Lanksheer, 1996). Second, 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 video games with friends and family is a large and normal part of the daily lives of youth. They further observed that playing video 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.
This proposal describes our plans to design, develop, and evaluate evidence-based assessments (embedded in a game) for a set of 21st Century competencies. When embedded assessments are so seamlessly woven into the fabric of the learning environment that they're invisible, we call this “stealth assessment” (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. During gameplay, students naturally produce rich sequences of actions while performing complex tasks, drawing on a variety of competencies. Evidence needed to assess the competencies is thus provided by the players' interactions with the game itself (i.e., the processes of play), which can be contrasted with the end product(s) of an activity—the norm in educational environments.

In this proposed research, we focus on developing and evaluating assessments for three 21st Century competencies: creativity, persistence, and understanding physics concepts in a well-designed digital game. The game we have selected is *Crayon Physics Deluxe*, a computer-based game that emphasizes two-dimensional physics simulations, including gravity, mass, kinetic energy, and transfer of momentum. Stealth assessments will collect ongoing and multifaceted information about the learner during gameplay while not disrupting engagement, and make reasoned inferences about competency states. Such competency estimates can also form the basis for diagnosis and instructional support, but that is beyond the scope of this current research proposal.

A key part of the assessment development process includes evidence-centered assessment design (ECD), which requires a systematic analysis of the assessment argument, including the claims to be made about the learner(s) and the evidence that supports those claims (Mislevy, Steinberg, & Almond, 2003). Information derived from gameplay (i.e., performance data) and associated activities (e.g., game-based discussions and help-seeking) will inform the competency estimates maintained within the student model. We will determine (a) the degree to which our stealth assessments yield valid and reliable measures of their respective competencies, and (b) if students learn conceptual physics from playing the game.

In addition to designing and developing stealth assessments and determining their reliability and validity, we also want to test how the stealth assessments may be scaled to other digital games that require similar competencies. This is important because the development costs of ECD-based assessments can be relatively high for complex competencies. Thus a second aim of this research is to establish a proof of concept for creating stealth assessment models that can be used in related games. Creating such cross-platform models for digital games would be useful and cost effective for educators interested in using games for assessment and support of learning. After we develop the stealth assessments in *Crayon Physics Deluxe*, we will analyze the efforts needed to use our stealth assessments effectively and efficiently within a different game (e.g., World of Goo, Swarm, or Super Meat Boy). The decision for our second game will be based on the degree to which the game can effectively elicit evidence of persistence, and the extent to which we can access the game code.
Past Work on Stealth Assessment

Currently, 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, in press). What no one has done yet is actually build stealth assessments directly within a digital game, as part of gameplay. Consequently, in this research we plan to build out three stealth assessments within a game to test: (a) the degree to which our stealth assessments yield valid and reliable measures of the respective competencies, (b) the effects of gameplay in relation to our selected competencies (e.g., improving understanding of conceptual physics), and (c) the ease (and challenges) of re-using our evidence-based models in a second game. We now describe the evidence-based design underlying stealth assessment.

Evidence Centered Design (ECD) (for a simple overview, see ECD for Dummies)

The primary purpose of an assessment is to collect information that will enable the assessor to make inferences about learners’ competency states—what they know, believe, can do, and to what degree. ECD defines a framework that consists of three main theoretical models that work in concert (see Mislevy & Haertel, 2006; Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, & Almond, 2003). The ECD framework allows an assessor to: (a) define the claims to be made about learners’ competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of problems that will elicit that evidence. A good assessment elicits behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence in terms that suit the purpose of the assessment. Thus the ECD approach is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design.

- **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.

In our proposed two-year research project, we focus on modeling and assessing three competencies: creativity, persistence, and physics understanding. Most of the work in Year 1 will involve developing and pilot testing the competency, evidence, and task models for our three competencies for use within our selected game (see Research Plan section for examples). Year 2 will focus on examining validity and reliability issues as well as learning during gameplay (focusing on physics understanding). We will also examine scalability issues related to our stealth assessments, discussed in the Year 2 sub-section of the Research Plan. Next is an overview of each competency that our three stealth assessments will be measuring in our digital game. The literature described below, coupled with additional research findings, will be used to fully develop the CMs used in our ECD-based assessments. The game will be discussed after the competencies.

21st Century Competencies

Creativity. Creative thought is a vital part of society and culture (Simonton, 1990; Walberg, & Stariha, 1992), playing a critical role in advancing science, technology, humanities, and the arts (Dudek, 2003). It has also become a key concern of organizations and businesses because of its role in innovation and entrepreneurship. Nearly 50 years ago, Bruner (1962) called for society to embrace and foster children's creativity as a preparation for the future since creativity comprises the foundation of problem solving. We define creativity as the ability to create novel solutions to various problems (Weisberg, 2003). Many problems in the world, and particularly in the sciences, are often complex, ill-defined, and cannot be solved unilaterally. Tomorrow's innovative thinkers need to be able to think "outside the box" using creative strategies to solve problems.

Persistence. Persistence can be broadly defined as the motivation to work hard despite challenging conditions (Peterson & Seligman, 2004). Persistence has consistently been found to predict academic achievement from preschool (Abe, 2005) to high school (Noftle & Robins, 2007, Proportat, 2009), to the postsecondary level (O’Conner & Paunonen, 2007) and adulthood (e.g., De Fruyt & Mervielde, 1996; Shiner, Masten, & Roberts, 2003). Meta-analyses have linked persistence with grades between $r = .21$ and $.27$, and the relationship between persistence and grades is independent of intelligence (e.g., Noftle & Robins, 2007, Proportat, 2009). Persistence holds many similarities to grit (Duckworth, Peterson, Matthews, & Kelly, 2007), defined as the combination of persistence and passion to attain long-term goals.

Conceptual Physics. Understanding physics concepts is key to understanding the physical world, and understanding Newton's Laws of Motion is foundational (e.g., diSessa, 1982; Savinainen, Scott, & Viiri, 2005), yet misconceptions are pervasive (e.g., Clements, 1991; Gunstone & Watts, 1985; Halloun & Hestenes, 1985). For instance, in their study of flawed "common-sense" conceptions of force and motion shared among college physics students, Halloun and Hestenes (1985) determined that only 15% of their 478 experimental subjects demonstrated an accurate
understanding of the relationship between unbalanced forces and acceleration (Newton’s 2nd law), while 61% confused the concepts of position, speed, and acceleration at least once.

**Competency Links to Common Core State Standards (CCSS)**

As described above, creativity and persistence have been shown to influence academic achievement and success in life. Moreover, having an understanding of physics can play an important role in developing strategies to deal with a wide range of problems in today's complex world (e.g., dealing with climate change and clean energy solutions). While our three focal competencies are not explicitly addressed in the CCSS, they can play a role in helping students learn the content embodied by the CCSS. In addition, there are some links between our competencies and the CCSS. For instance, here is an excerpt from the CCSS regarding creativity in high school mathematics: "Real-world situations are not organized and labeled for analysis; formulating tractable models, representing such models, and analyzing them is appropriately a creative process. Like every such process, this depends on acquired expertise as well as creativity" (p. 72).

In relation to mathematical practices, the very first standard involves persistence: "Make sense of problems and persevere in solving them" (p. 6). Finally, improving students' understanding of physics concepts may transfer to various areas in mathematics, such as geometry and algebra. One particular standard from the CCSS that is related to conceptual physics is mathematical modeling: "Modeling is the process of choosing and using appropriate mathematics and statistics to analyze empirical situations, to understand them better, and to improve decisions. Quantities and their relationships in physical... situations can be modeled using mathematical and statistical methods" (p. 72). Thus improving conceptual physics understanding could transfer to problem solving in more formal mathematics courses that require modeling.

The next section introduces the digital game we will be using as a vehicle to assess our three competencies.

**Research Plan**

The game we have selected to use for our research is Crayon Physics Deluxe. Using this game, we aim to answer the question: *To what degree are the stealth assessments we develop valid and reliable measures of the respective competencies?* Below is a description of the game and examples of evidence indicators (i.e., student behaviors) that can be obtained from gameplay to inform our competency models.

*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," the player must draw a golf club on a pulley wheel to make it swing to hit the ball. The speed of the swinging golf club is dependent on the size/mass 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. The game designer and developer (Petri Purho) has agreed to participate in this project with us.

Various problems in CPD require the player to use catapults, pulleys, and so forth to move the ball. A screen capture is shown on the left, with the ball on the left, the star on the right, and some gears/chains obstructing the way. Solutions are drawn with crayons using the mouse. Other problems require players to draw unique shapes that can be used to move the ball towards the star. 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. To illustrate, as we (PIs on the project) spent time "examining" the game in depth prior to selecting it for our research vehicle, it was not uncommon for each of us to revisit/replay particularly challenging problems at least ten times, always striving for a better, more elegant solution.

Evidence for creativity in CPD. All of the problems in the game have multiple solutions which can be exploited to assess variation in problem solving. This variation in solutions will comprise the evidence models in our creativity stealth assessment. For example, one problem in CPD requires the player to guide a ball down a series of platforms to reach the star. The player can create blockades, a series of ramps (or one long ramp), swinging pulleys, and/or weights to guide the ball down the platforms without the ball falling into a pit. Thus players have the freedom to create their own combination of steps to reach a solution. Furthermore, the game automatically assigns varying points for different types of solutions. For instance, working out a solution to a problem using a single object is usually very hard and thus is judged to be an "elegant" solution giving the player an extra point. The number of elegant solutions (relative to other solutions) could be one good indicator of creativity.

Evidence for persistence in CPD. Persistence is relevant to gaming environments because learning how to master an aspect of a game or solve a particularly wicked problem can be difficult and requires the will to work hard despite repeated failure (Peterson & Seligman, 2004). Being willing to work hard is a core disposition that we believe is important for 21st Century success and can be assessed rather easily in games. Solving problems in CPD can be quite challenging, and a player may try dozens of approaches to solving a problem before reaching a successful solution (note: there is no explicit help offered to players in CPD). In some cases the player might just need to skip the problem and come back to it later. Some of the potential observables for persistence include: number of attempts at solving each problem, time spent on a given problem, how often a player who skipped a problem returned to it later, and so on. Much of this information is already captured by the game.

Evidence for physics understanding in CPD. Physics engines are becoming pervasive in gaming environments, providing a sense of realism in a game (e.g., Havoc 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. We plan to use many of the scoring features built into the game to create our evidence models for physics understanding. For instance, like the "elegant" solution described above, a solution using only basic principles of physics (with no nudges of the ball or fancy gadgets) is classified by the system as "old school" and gives the player an extra point.

Building competency estimates from evidence indicators. Evidence indicators like the ones just described will be used to dynamically construct probabilistic competency estimates for a student (e.g., "The probability that Emily is at the highest level of proficiency for Creativity is .94"). Evidence indicators will be weighted and updated in a Bayes Net in each EM. Over time, students with an accumulation of positive evidences of the competency (i.e., as defined in the EM) will have higher competency estimates. Each competency will have dozens of evidence indicators.

Figure 1 displays an example of how we could link a simplified creativity competency model to evidence indicators derived from CPD gameplay. For instance, creativity may be decomposed into two broad facets (i.e., novelty and fluency) which are linked to evidence indicators in CPD. The first indicator for novelty, "novel use of objects," is a behavior that is determined by the number of times a player creates and/or uses objects in uncommon ways (i.e., as defined by the sample) to solve a problem. The evidence indicator value can then be normalized by the number of times the player (or sample of players) executes this behavior. Based on pilot work, we will be able to apply differential weights to evidence indicators to represent difficulty parameters in the model. These difficulty parameters could be decided based on theory in our competency model, and by data based on student performance across the evidence indicators.

Figure 1. Illustrative Competency Model for Creativity With Evidence Indicators

The next section describes our general plans for developing the full set of ECD models.

Year 1: Designing ECD Models and Piloting CPD Problems

Making use of ongoing streams of evidence to assess students' competencies 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 can affect subsequent actions later on. Second, in traditional tests, questions are often designed to get at one particular piece 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 problem (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.

To address these issues, the general activities in Year 1 will involve creating the three competency models and the associated evidence models (with relevant indicators). To define parameters of our task model we will use existing problems in CPD as well as create new ones with the game's "level editor" tool. As part of developing the task models we will identify problems that most effectively elicit evidence related to the three focal competencies. We also plan to pilot test the problems—existing and created ones—to determine if they're appropriate for our population and our methodological requirements (e.g., adequate variability among the problems). Pilot work will be conducted on middle school students at Florida State University School (FSUS; where we have permission by the school to conduct this research) and possibly the Florida Virtual School (FLVS), with whom we plan to initiate contact directly after the proposal is funded.

**Year 2: Validity and Scalability Research**

Two studies will be conducted to evaluate the validity and reliability of the stealth assessments, and the scalability of the developed CMs and EMs. Study 1 (n = 120) will evaluate the validity and reliability of our three stealth assessments in CPD. At the beginning of the study, students will complete three traditional tests on our focal competencies (e.g., Torrance test for creativity; relevant items from the International Personality Item Pool for persistence; and Force Concept Inventory for physics, one of the most widely-used and validated assessment instruments in physics education, Hestenes & Halloun, 1995; Hestenes, Wells, & Swackhamer, 1992). We will also collect GPA information on students. Next, students will interact with 15 CPD problems over three one-hour sessions in the computer lab at their school (i.e., roughly five problems each hour). After each hour-long session, the three competency levels will be estimated from students' gameplay in CPD.

At the conclusion of the last CPD session, students will complete an isomorphic posttest of physics understanding that will be compared to the physics pretest given at the beginning of the study. In addition, final competency estimates1 will be correlated with the scores obtained from the traditional tests of the same competencies. The results of the study will inform us as to the validity of the stealth assessments for the three focal competencies (creativity, persistence, and physics) and provide us with preliminary evidence for physics learning in CPD. Given that understanding conceptual physics is not subject to natural development or maturation, examining pretest-posttest differences will suffice to determine learning as a function of gameplay.

---

1 To reduce the probability estimates (e.g., high, medium, and low levels) to a single number, we plan to assign numeric values +1, 0 and -1 to the three competency states, and compute the expected value. This Expected A Posteriori (EAP) value can also be written as, $P(θ_i = \text{High}) - P(θ_i = \text{Low})$, where $θ_i$ is the value for Student $i$ on Competency $j$, and $1*P(\text{High}) + 0*P(\text{Med}) + -1*P(\text{Low}) = P(\text{High}) - P(\text{Low})$. This results in a scale ranging from -1 to 1.
During the second year, we also plan to analyze the degree to which we can "plug 'n play" our stealth assessment models, from one game to another. That is, we will attempt to reuse the same persistence competency model and associated evidence models developed in Year 1 in a different digital game, such as the World of Goo. In theory, we should only have to modify the task models for persistence as the competency model and the majority of the evidence models should be applicable to the second game. For example, one persistence indicator that is shared between World of Goo and CPD is the player behavior of "returning to skipped problems and eventually solving them." This provides evidence of persistence as it shows the extent to which players persist despite failure. Thus the number of evidence indicators that the second game shares with CPD will be the main factor in deciding on the second game. After deciding on the second game, we will modify the task models. This entails specifying and identifying problem types in the second game that can be used to elicit the same evidence for persistence as is specified in CPD (e.g., time spent trying to solve a problem, replaying a problem already solved to improve the score, and so on).

A subset of the students from Study 1 will be used in Study 2 \((n = 80)\) to evaluate the effectiveness of the recycled stealth assessment. Students will interact with the second game over two one-hour sessions in the computer lab. During each session, estimates of persistence will be updated based on student gameplay. After the sessions, the Study 2 persistence estimates will be correlated to the Study 1 persistence estimates in CDP, and to scores from the traditional tests. This will provide information about the validity of the second-game assessment implemented in Study 2. Scalability (i.e., cost-benefit analysis) will be examined relative to the challenges (costs) associated with the time it takes to modify the TM to suit the second game, and the effort needed to insert the CM and EMs into the second game. We will monitor costs in terms of the time required for these efforts, and examine costs relative to the validity of the competency estimates.

**Data Analyses**

*Reliability and validity of stealth assessments.* Split-half reliabilities will be calculated across all evidence indicators in CPD for each competency. Thus if there are, say, 40 evidence indicators of creativity in CPD, we will split those indicators into two data sets (i.e., compiled over all students) and correlate the competency estimates between the two sets. We expect to see a strong correlation among indicators for each competency \((r \geq .80)\). We will evaluate construct validity of stealth assessments by computing correlations between the competency estimates and the matched traditional measures. We expect to see moderate positive correlations between the stealth assessment competency estimates and the traditional assessment scores \((r = .40 - .60)\). We will also evaluate the predictive validity of the stealth assessments by computing hierarchical regression analyses to investigate how well the stealth assessment competency estimates predict student GPA over and above the predictive ability of the traditional measures. We expect the stealth assessments to predict GPA beyond the traditional measures. Finally, we will run confirmatory factor analyses to evaluate the independence of each of the three competencies for both the stealth assessments and traditional assessments. We expect to see a better model fit (i.e., stronger evidence for independence among the three competencies) for the three stealth assessments compared to the three traditional assessments.
Learning gains. We will compute a paired sample t-test to compare physics learning before playing CPD versus after playing CPD. We expect to see a significant difference between pretest and posttest. Covariate measures (e.g., gender, grade) will be used to control for any individual differences variance.

Reliability and validity of persistence stealth assessment in second game. We will compute split-half reliabilities across all evidence indicators for persistence in the second game. We expect to see a strong correlation among indicators for persistence (e.g., $r \geq .80$). Then we will compute correlations among the persistence estimates in Study 1 and Study 2 to examine test-retest reliability. We expect to see a strong correlation between the persistence estimates calculated in Study 1 and Study 2 ($r \geq .60$). Additionally, we expect to see the persistence estimates in Study 2 predict student GPA because persistence measures have been shown to predict GPA (see p. 5).

Summary

Our country's current approach to assessing students has a lot of room for improvement, at the classroom as well as the high-stakes levels. This is especially true in terms of the lack of support that current/traditional 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. Our proposed solution to this problem is stealth assessment. The concept of stealth assessment is garnering increased attention over the past couple of years. For example, if you Google the term "stealth assessment," you'll see a number of references to our idea (and to see a featured article in the Chronicle of Higher Education on the topic, click: http://chronicle.com/article/A-Stealth-Assessment-Turns/125276/). We have developed a robust methodology as well as several substantive mockups for various games, but it's time to develop and evaluate stealth assessments within actual game environments, which comprises the crux of this research proposal.

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 our trio of 21st Century competencies, then the next logical step is to examine scalability – what are the costs and benefits of reusing ECD-based models in different games to assess the same kinds of competencies?

Alignment of Strategy

The purpose of this grant proposal is in line with the strategic goals of the Gates Foundation. For instance, the Gates and MacArthur Foundations recently funded the Games, Learning, and Assessment Workshop (January 27-28, 2011) aimed to facilitate discussion about how to use digital games as tools for assessment. The workshop invited assessment and learning experts, school principals, district leaders, and game-industry persons (designers, developers, and user researchers) to discuss how best to create a research process that uses principles of game design and contemporary learning theories to suit the needs of an assessment system. The first goal of the workshop was to see how feasible it is to combine assessment and game design expertise towards the design of new game-based assessments. A second goal of the workshop focused on
putting assessment at the center of the discussion around games and learning rather than as an after-thought. The third goal of the workshop involved examining issues related to using games as assessment vehicles under the Common Core State Standards.

This proposal addresses the goals of the workshop in three main ways. First, we will address the extent to which ECD-based assessments can be implemented in digital games, and further establish the validity and reliability of our stealth assessments. Second, we will evaluate the ease (and challenges) of implementing stealth assessment models within multiple games. This would provide a cost-effective way for educators and researchers to systematically explore how games can be used for assessment and learning. Finally, the focal competencies of interest seem to be aligned with the CCSS, both generally in terms of the goals of the CCSS, and in relation to particular standards.

**Organizational Capacity**

An interdisciplinary team has been assembled to achieve the goals of this project. Dr. Valerie Shute (PI) is a Professor in the Educational Psychology and Learning Systems department at Florida State University (FSU). She is an educational psychologist, designer of numerous systems to promote learning, and an expert in diagnostic assessment. She will direct the entire project. Dr. Matthew Ventura (Co-PI) is a visiting professor at Florida State University. He is a cognitive scientist with expertise in educational technology, natural language understanding, and noncognitive assessment. Before coming to FSU Matthew was the research director of the ETS Personal Potential Index (PPI), the first noncognitive assessment to play a role in high stakes admission decisions in graduate school. The PPI is a rating form for mentors (e.g., faculty or employers) to rate students on five 21st Century competencies (e.g., knowledge and creativity, teamwork, resilience, planning and organization, ethics and integrity) for graduate school admissions. Matthew’s primary responsibility will be to oversee the design of competency, evidence, and task models for persistence and creativity. Dr. Russell Almond, a psychometrician in the Measurement and Statistics program at Florida State University, as well as one of the original founders of ECD, will be overseeing the statistical training algorithms in all the ECD models. We will also hire a physics instructor at FSU who will help develop the conceptual physics competency and evidence models. Additionally, we will hire a computer science consultant with a background in artificial intelligence techniques responsible for abstracting log information from CPD for our ECD models. Finally, Petri Purho, designer and developer of CPD will provide support for any game modifications that are needed to aid the assessment development process.

Regarding our team’s ability to perform work of this type and scope, in the past five years, the PI (Shute) has received more than $3,500,000 in grant awards (where, on 7/9 of the awards, she was PI). Two members of our team also recently received a U.S. patent for advanced technology relevant to this proposed research (i.e., *Method and System for Designing Adaptive, Diagnostic Assessments*, #US 7,828,552; Shute, V. J., Hansen, E. G., & Almond, R. G., awarded November 9, 2010). Furthermore, subsets of our team have worked on various research projects in the past. For instance, Shute and Almond collaborated on a recent three-year NSF grant (Shute (PI), Adaptive e-learning for Middle School Mathematics. Funded by National Science Foundation (NSF-REC #0313202). And Shute and Ventura teamed up to write the first paper on stealth assessment (Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D., 2009. *Melding the
power of serious games and embedded assessment to monitor and foster learning: Flow and grow). In the past five years, we have co-authored dozens of journal articles, book chapters, and even a recent report for the Gates Foundation (i.e., Shute, V. J., Ventura, M., & Kim, Y. J., 2011. Synthesis report on the games, learning, and assessment (GLA) Workshop).

Regarding our targeted school’s capacity to support this research, the Florida State University School (FSUS) has a diverse student population; approximately 56% of the students are minority status (including 31% Hispanic and 25% African American) and 45.2% come from low-income families (Florida DPE Series 2010-22D, 2010). Thus our target population will be a heterogeneous sample of middle school students (i.e., grades 6-8; 11–14 years old). All studies will be conducted at FSUS. Students, along with a group of science teachers, will be recruited from the participating school. Participant recruiting will continue throughout the two-year program with assistance of the school administration. We have begun the process of IRB approval for the two-year study and have obtained permission from FSUS for testing. The PI will be responsible for ensuring that all of Florida State University’s IRB policies and procedures are followed for the duration of the project. We also plan to initiate contact with the Florida Virtual School (FLVS) for additional testing of the stealth assessment research herein. FLVS provides virtual K-12 education solutions to students all over Florida, the U.S. and the world. They were founded in 1997, are nationally recognized for their e-learning model, and have received numerous awards.

**Risks**

This research is exploratory and runs the risk of yielding little viable use if null results are found. Additionally, there may be factors outside of our control that may impede the completion of the project. This would include obstacles in obtaining a sufficient sample at FSUS (which is not likely to be an issue).

**Measurement Learning and Evaluation**

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). Well-designed digital games can provide meaningful assessment environments by providing students with problems that require the application of various competencies.

This research will evaluate the reliability, validity, effectiveness, efficiency, and scalability of our stealth assessments. Three stealth assessment will be created based on gameplay in CPD and evaluated for their reliability and validity. If the stealth assessments built in CPD are valid and reliable, we will evaluate the extent to which the persistence stealth assessment can be modified and validated in a different digital game. We will publish several journal articles on this research related to (a) the reliability and validity of stealth assessments in CPD, and (b) the processes involved with scaling evidence-based stealth assessment models to a different digital game.
Sustainability

This research will permit us to rigorously evaluate the validity and reliability of the stealth assessment approach within a game. It will also let us evaluate the challenges and costs associated with reusing the ECD models from one of the stealth assessments (i.e., persistence) within another game. If this particular evidence-based stealth assessment methodology is found to be valid and reliable, we plan to make the process and models broadly available so that the work will continue. One idea for dissemination includes posting the results and models on the workingexamples.org web site for others to view and use. We can also disseminate via the new "games, learning, and assessment" research area that has emerged from the recent Gates-MacArthur Workshop on the topic. And we plan to publish our findings in peer-reviewed journals and make our models available to other researchers.

If this Phase 1 research is successful, Phase 2 (follow-on) research can expand in a number of directions. First, we 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 CCSS (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 to determine the range of games that may employ the same competency and evidence models, for a scalable, cost effective and engaging solution to assessment of complex competencies. Finally, stealth assessment 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).

Sustaining this research is critical to us as our research is centered around games, learning, and assessment. We will continue to seek funding to advance the field and publish findings in scholarly journals. We will also create supporting documentation on how to apply the stealth assessment models to other digital games that support similar competencies.

References


Florida State University


