Stealth Assessment in Portal 2

Project summary

This project will investigate the video game Portal 2 (by Valve Corporation) as a vehicle to assess and support competencies important for success in today's world. Portal 2 can be classified as a first-person, puzzle-platform, "shooter" game, consisting of a series of puzzles. These puzzles must be solved by teleporting the player's character and simple objects using the "portal gun," a device that can create inter-spatial portals between two flat planes. To solve the progressively more difficult challenges players must figure out how to locate, obtain, and then combine various objects effectively to open doors and navigate through the environment.

In Year 1, we will work with Valve to develop stealth assessments for three target competencies—problem solving skill, spatial ability, and persistence. In Year 2 we will evaluate the validity of the stealth assessments and examine any learning that occurs as a function of playing the game. High school students will play Portal 2 for eight hours (broken up into 5-10 gaming sessions). During each session, stealth assessments will record data extracted from the gameplay log files, and competency estimates will be updated in real time to reflect changes in competency states. At the beginning and end of the Portal 2 sessions (i.e., before the first session and after the final one), students will take traditional assessments of our selected competencies. Scores from the traditional measures will be compared to the competency estimates obtained through the stealth assessments collected over time spent playing Portal 2 to establish validity. To measure learning for each of the three competencies, we will compare pretest - posttest gains on the traditional measures from the Portal 2 group with those of a control group (*n* = 50). We will also analyze growth in the estimates of competencies obtained from our stealth assessments.

Problems we want to address

We will address several issues with this research. First, despite increased interest in and discussion of game-based learning, little empirical research has examined games' effects on learning (e.g., Charsky, 2010; Van Eck, 2009). That is, compared to other types of instructional and assessment systems, few experimental studies have examined the range of effects of gaming environments on learning. Similarly we found a corresponding lack of theory and practice for the design and implementation of such systems.

The second issue we want to tackle concerns the problems associated with improving learning and assessment in K-12. Teachers need to move beyond a simple content-learning mindset and towards assessing and supporting important skills for success in the 21st Century (Gee, 2005). Additionally, we need to focus more on the use of assessment formats that can help to improve learning. For example, in our current, complex world, being able to solve ill-structured problems is, and will continue to be, of utmost importance. We are confronted daily with "wicked" problems of enormous complexity and global ramifications (e.g., the dysfunction in Congress leading to a downgraded rating by S&P, global warming, destruction of the rain forests, and so on). The people who will be making and managing policy decisions in the near future need to be able to understand, at the very least, how research works and how science works because solutions are going to be incredibly technical and highly complex. When confronted by such problems, the ability to think creatively, systemically, and not give up when the going gets rough is essential. Learning and succeeding in a complex and dynamic world is not easily or optimally measured by multiple-choice responses on a simple knowledge test. Instead, solutions to problems begin with re-thinking assessment, identifying skills relevant for the 21st century, and then figuring out how best to assess students' acquisition of the new competencies.

We now frame our questions relevant to this research proposal.

Research questions

We aim to answer two main questions in this two-year project:

1) *Validity*—Do the stealth assessments measure what we think they are measuring? That is, to what extent are the stealth assessments valid and reliable measures of our target competencies—problem solving, spatial abilities, and persistence?

2) *Learning*—To what extent do the target competencies improve over time by playing Portal 2? We will investigate learning in three ways: (a) internally in the game via growth in stealth assessment competency estimates, (b) externally via pretest - posttest differences on traditional assessments (transfer), and (c) comparatively against a control group.

Learning in Portal 2: Practice makes perfect

People who want to excel at something—from athletes to dancers to surgeons to video game programmers—spend countless hours practicing their craft. By continually refining techniques and developing new maneuvers to enhance their skills, they manifest the belief that practice is critical to improvement. There's considerable support in the literature, going back more than 100 years, for the idea that "practice makes perfect," or in its less extreme form, that "practice makes better" (Bryan & Harter, 1899; Newell & Rosenbloom, 1981; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977; Shute, Gawlick, & Gluck, 1998; Underdahl, Palacio-Cayetano, & Stevens, 2001).

The historical foundation for contemporary research on this topic was established by Thorndike's (1898) investigations into the effects of practice with feedback. The common conclusion across all of this work, oversimplified, is that people become more accurate and faster the more often they perform a task. Learning or skill acquisition thus represents a change in a person that occurs as a function of experience or practice. Because it's not directly observable, learning must be inferred from performance on some task. But practice can be boring and frustrating, causing some learners to abandon their practice and hence learning. This is where good digital games come in – they can provide an engaging environment designed to keep practice interesting.

Practice has been operationalized through the term "deliberate practice," which is defined as engaging in a training activity designed to improve some aspect of performance with full concentration, analysis after immediate feedback, and opportunities for repetition with refinement (Ericsson, 2006; Ericsson & Ward, 2007). Deliberate practice involves working at the edge of one's abilities. Similarly, Vygotskian theory (e.g., Vygotsky, 1987) proposes that effective learning occurs in the "zone of proximal development," where the child is able to perform beyond her own ability with the help of adults or with other means of support (e.g., automated feedback). By presenting kids with these types of challenges (i.e., at the very outer limits of their skill level), the Vygotskian framework has been shown, in random-assignment studies, to improve learning in disadvantaged children (Barnett et al., 2008; Bodrova & Leong, 2007; Diamond, Barnett, Thomas, & Munro, 2007).

Can problem solving skill be improved with practice? Polya (1945) has argued that problem solving is not an innate skill, but rather something that *can* be developed, "Solving problems is a practical skill, let us say, like swimming… Trying to solve problems, you have to observe and imitate what other people do when solving problems; and, finally, you learn to solve problems by doing them." (p. 5). Students are not born with problem solving skills. Instead, these skills are cultivated when students have opportunities to solve problems proportionate to their knowledge.

According to Jonassen (2002), all good problems share two characteristics. First, they have some kind of goal, or unknown. The goal/unknown requires the generation of new knowledge. Second, all problems should have some value to the learner in solving them. Games similarly have a set of goals and unknowns which require the learner to generate new knowledge. Games (i.e., good ones like Portal 2) also have value to the learner in terms of achieving the challenging goals. For more details on problem solving, see the section below on *Target competencies in Portal 2*.

Assessment in Portal 2: Engagement is paramount

Just like in learning environments, assessments can be deficient or invalid if the tasks or problems are not engaging, meaningful, or contextualized. This need for more authentic and engaging assessments has motivated our recent research efforts to re-think assessment, particularly as it can occur naturally within good games. In contrast, the amount of engagement with paper and pencil, multiple-choice assessments can be negligible. When these problems associated with traditional assessment—inauthentic and decontextualized items, and anxiety—are removed (e.g., by using a well-designed game as the assessment vehicle), then the assessment should be more engaging. Additionally, if the assessment is designed properly, such as by using an evidence-centered design approach (Mislevy, Steinberg, & Almond, 2003), then it should be as (or more) valid compared to a traditional assessment.

Using games as assessment vehicles has its own set of issues. For instance, video game assessments have potential sources of error variance such as varying levels of interest in the target game. However, we believe this will not be a major problem with Portal 2 given its broad appeal (i.e., over 3 million copies have been sold since it came out last year, according to GameFront). In short, we believe that Portal 2 can be used to assess competencies better than traditional assessments, by virtue of having more authentic, contextualized, and engaging tasks. Below are the competencies that we believe are required to succeed in Portal 2 and thus should be assessed in the game.

Target competencies in Portal 2

Problem solving. The development of problem-solving ability has often been regarded as a primary goal of the education process (Ruscio & Amabile, 1999). Although schools traditionally have advocated the instruction of basic content such as reading, writing, and mathematics (Glaser, Pellegrino, & Lesgold, 1978), promoting problem solving emerged in the 1980s as a way to facilitate general thinking and reasoning skills (Bransford, Arbitman-Smith, Stein, & Vye, 1985).

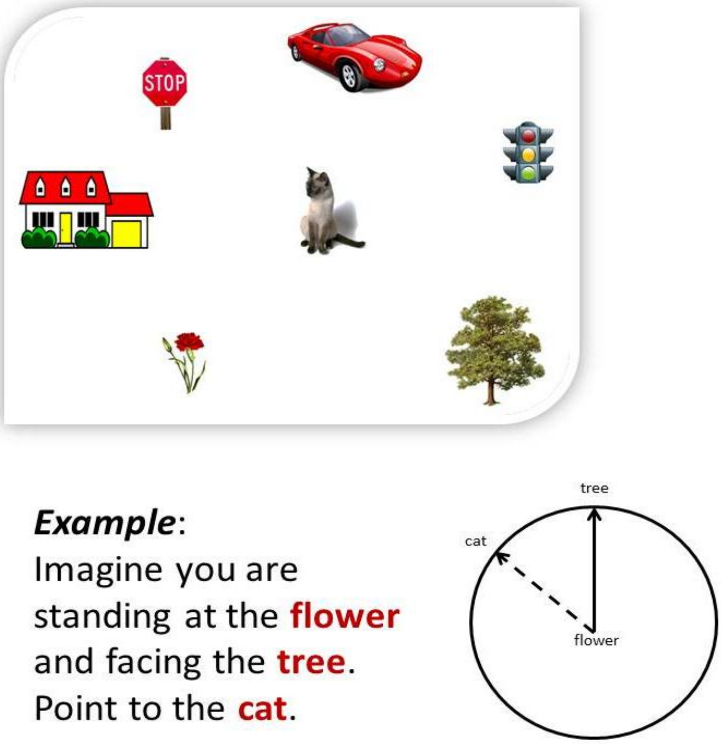
Psychologists and educators typically describe problem solving as involving a cycle that includes (a) problem identification, (b) problem representation, (c) hypothesis generation, (d) hypothesis testing, (e) progress monitoring, and finally (f) evaluation of the implemented solution (see, for example, Bransford & Stein, 1993; Hayes, 1989; Sternberg, 1999). One cognitive process thought to hinder problem solving is functional fixedness, defined as the difficulty that a person experiences when attempting to think about and use objects (or strategies) in unconventional ways (Duncker, 1945). This cognitive rigidity causes people to view a particular type of problem as having one specific kind of solution without allowing for alternative strategies and explanations (Anderson, 1983). Many problem-solving strategies that are taught in school entail a "cookbook" type of memorization, resulting in functional fixedness which can obstruct students' ability to solve problems for which they have not been specifically trained. Additionally, this pedagogy also stunts students' epistemological development, preventing them from developing their own knowledge-seeking skills (Jonassen, Marra, & Palmer, 2004).

One direct effect of playing Portal 2 could be the reduction in functional fixedness. As pointed out earlier, Portal 2 provides a unique gameplay environment that can promote problem solving skills through providing players extensive practice figuring out solutions to problems on their own. This practice reinforces players to rethink solutions to problems and be flexible in terms of trying alternative strategies. This is important since the way in which students learn problem-solving strategies may influence their subsequent ability to understand and flexibly apply this information in the world (Bransford, Vye, Kinzer, & Risko, 1990; Ruscio & Amabile, 1999; Voss, 1989)

Spatial abilities. Spatial abilities generally refer to two broad skills consisting of large-scale spatial cognition (i.e., orientation), and small-scale or object-based spatial cognition (e.g., mental rotation tasks; Hegarty, Crookes, Dara-Abrams, & Shipley; in press). Large-scale spatial cognition refers to one's sense of direction or the ability to navigate through the environment effectively and efficiently (Darken, Allard, & Achille, 1998; Shute, 1984). Findings have shown that the hippocampus plays a significant role in maintaining what are called "place cells" which help people maintain a mental map of their surroundings for navigation through new spaces. Studies have shown that the hippocampus is actually larger in taxi drivers versus non taxi drivers (Maguire, Frackowiak, & Frith, 1997) which provides evidence that spatial abilities can continue to grow throughout the lifetime.

Spatial abilities have also been studied across professionals in various scientific disciplines (Hegarty, Crookes, Dara-Abrams, & Shipley, in press). Those successful in the hard sciences (e.g., geology, engineering, and chemistry) seem to show higher levels of both spatial orientation and mental rotation skills than those in the soft sciences (e.g., psychology, humanities, and social sciences). Finally, sense of direction and mental rotation skills show a low correlation ($r < .3$) suggesting that these skills are relatively independent and thus could be studied separately (Hegarty, Montello, Richardson, Ishikawa, & Lovelace, 2006).

While spatial ability assessments have been around for decades, 3-D spatial orientation assessments are fairly new and require sophisticated simulation equipment. One such test is the 3-D Perspective taking test (Kozhevnikov, 2008). In this test, a person wears a special helmet that displays a simulated virtual environment. The person is then asked questions about the location of various objects and must physically point to objects in the 3-D environment. A simpler test assessing spatial orientation is the 2-D Spatial Orientation Test (Hegarty & Waller, 2004). In this test, a person must judge the direction of objects on a 2-D plane. This test assesses a person's sense of direction by asking where certain objects would be located if one was facing a particular direction. See Figure 1 for an example.



Example:
Imagine you are standing at the **flower** and facing the **tree**. Point to the **cat**.

Figure 1. Spatial Orientation Test (Hegarty & Waller, 2004).

Empirical studies have shown mixed results regarding success in improving spatial abilities (e.g., Anderson & Bavalier, 2011; Mayer, 2011; Pellegrino, 1984). Portal 2 provides a gameplay experience that forces players to use their spatial abilities to solve problems. Practicing with multiple and varied spatial problems over time should lead to improvement in spatial abilities.

Persistence. Persistence can be broadly defined as the motivation to work hard despite challenging conditions (Peterson & Seligman, 2004). Persistence is considered to be a facet or sub-category of conscientiousness (MacCann, Duckworth, & Roberts, 2009), a disposition which has consistently been found to predict academic achievement from preschool (Abe, 2005) to high school (e.g., Noftle & Robins, 2007, Proporat, 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 conscientiousness with grades, with mean correlations between $r = .21$ and .27, and the relationship between conscientiousness and grades appears to be independent of intelligence (e.g., Noftle & Robins, 2007; Proporat, 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.

Persistence measures, like most dispositional measures, are primarily self-report (e.g., *I work hard no matter how difficult the task*), a method of assessment that is riddled with problems. First, they 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. Portal 2 shows promise as a vehicle to assess persistence using performance-based assessments of dispositional competencies. These performance-based assessments can record and score actual behaviors in Portal 2 that pertain to a particular competency. Additionally, Portal 2 provides a gameplay environment that can improve persistence, because many problems require players to persevere despite failure and frustration. Portal 2 can be quite difficult, and we believe that 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.

Assessing competencies in Portal 2. Working with Valve will enable us to identify many observables in gameplay that will be used to inform our target competencies. For example, problem solving skills can be assessed in the game by evaluating the actions and behaviors one engages in during Portal 2 gameplay. To illustrate, a typical problem requires the player to use a portal gun and various objects (e.g., boxes, lasers, laser refractors, trampolines, buttons) to open a door in a chamber (large room or set of connected rooms). Players must figure out how to correctly combine the various objects (e.g., use the portal gun to "teleport" to a platform, then use a box to climb up to another platform to turn on a laser that activates a sensor to open a door) in order to complete a problem. Spatial orientation can be assessed in Portal 2 by observing how often players get disoriented in gameplay. That is, certain patterns, like moving around in circles within the same space, yield evidence that a player is spatially disoriented. Other patterns may demonstrate good orienting skills—such as quickly scanning the space upon landing out of a portal to get one's bearings. Persistence may be assessed in Portal 2 by seeing how often and how long players continue to play a problem despite repeated failure.

Interdependencies among competencies. We expect to see some interdependencies among the three competencies. For example, problem solving in the game will likely require spatial abilities and persistence. Without spatial orientation and persistence, one could not adequately represent the problem space in Portal 2 nor make it to the end of the game.

Year 1: Designing ECD Models and Piloting Portal 2 Problems

The primary purpose of any 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. Accurate inferences of competency states support instructional decisions that can promote learning. Evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003) defines a framework that consists of three theoretical models that work in concert. The ECD framework allows/requires 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 tasks that will elicit that evidence. These three actions map directly onto the three main models of ECD shown in Figure 2.

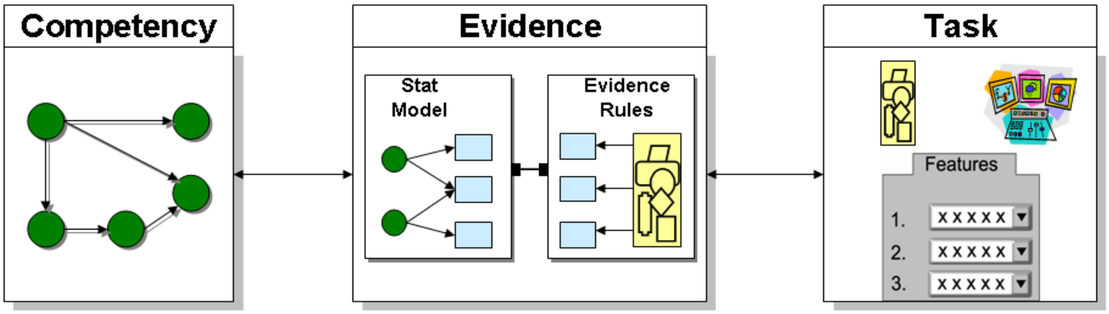


Figure 2. Three main models of an evidence-centered assessment design

A good assessment has to elicit 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. Working out these variables, models, and their interrelationships 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? This can also be phrased as: What do you want to say about the person at the end of the assessment? Variables in the competency model (CM) are usually called "nodes" and describe the set of person variables on which inferences are based. The term student (or learner) 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 learner model express the assessor's current belief about the level on each variable within the learner's CM.

- *Evidence Model*: What behaviors or performances should reveal those constructs? 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 connection between those behaviors and the CM variable(s)? Basically, an evidence model lays out the argument about why and how observations in a given task situation (i.e., student performance data) constitute evidence about CM variables.

- *Task Model*: What tasks (e.g., problems in the game) should elicit those behaviors that comprise the evidence? A task model (TM) provides a framework for characterizing, constructing or identifying situations with which a learner will interact to provide evidence about targeted aspects of knowledge or skill related to competencies. Task specifications establish what the learner will be asked to do, what kinds of responses are permitted, what types of formats are available, and other considerations, such as whether the learner will be timed. Multiple task models can be employed in a given assessment. Tasks are the most obvious part of an assessment, and their main purpose is to elicit evidence (which is observable) about competencies (which are unobservable).

As shown in Figure 2, assessment design flows from left to right, although in practice it's more iterative. Diagnosis (or inference) flows in the opposite direction. That is, an assessment is administered, and the learners' responses made during the solution process provide the evidence that is analyzed by the evidence model. The results of this analysis are data (e.g., scores) that are passed on to the competency model, which in turn updates the claims about relevant competencies (i.e., probabilities). In short, the ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about personal competencies via an evidentiary chain (e.g., valid arguments that serve to connect task performance to competency estimates), and are thus valid for their intended purposes.

The general activities in Year 1 will involve creating the competency models and the associated evidence models (with relevant gameplay indicators) for our three main competencies: problem solving skills, spatial abilities, and persistence. To define our task model we will use existing problems in Portal 2 and possibly create new ones with the game's "mod" tool. As part of developing the task models, we will identify problems that effectively elicit evidence related to the competencies. Problems will also be arrayed by difficulty to provide for adaptive challenges to support the "zone of proximal development" (Vygotsky, 1987) and "flow" state (Csikszentmilhalyi, 1990). Valve Corp. will work with us to (a) modify the source code to enable the abstraction of relevant data for our stealth assessments, and (b) help us set up specific sequences of problems for testing.

We plan to pilot test select Portal 2 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 likely be conducted at Valve Corp.

Year 2: Evaluation and Assessment

One large study will be conducted in Year 2 to evaluate the validity and reliability of the stealth assessments, as well as learning as a function of gameplay in Portal 2. At the beginning of the study, students will complete traditional tests on our target competencies (i.e., validated measures for problem solving, spatial abilities, and persistence). We will also collect GPA and FCAT information on students. Next, students will interact with Portal 2 over eight hours (spaced across 5-10 gaming sessions spanning 2-3 weeks) in the computer lab. We anticipate running the study with high school students (e.g., 10th-12th grade) in a computer lab at FSU.

Design

To answer our two research questions related to validity and learning, around 230 students (total), across grades 10-12 will be needed. In terms of our learning evaluation, we will run 180 students in the Portal 2 group and 50 in the control group (note: due to the difficulty in running 10-hour studies, we had to reduce the sample size of the control group, but $n = 50$ is still within the acceptable limits to detect significant differences between the treatment and control group; Faul, Erdfelder, Buchner, & Lang, 2009). Additionally, 180 students is a sufficient sample size to test the various reliability and validity questions regarding the three stealth assessments and their corresponding traditional assessment scores. Students will be randomly assigned to one of two different conditions.

Conditions

1) *Control group*—50 students will be in our control group. They will play various casual video games for 2-3 weeks (8 hours, total). The games (e.g., Sonic the Hedgehog, Pac-Man, Space Invaders) will be played in flash-based versions on the computer. These games are intended to serve as an activity that is engaging (and comparable to the treatment group), but not heavily dependent on our three target competencies. This group will provide (a) baseline data concerning pretest-posttest differences, and (b) data to permit us to make causal inferences regarding competency growth in Portal 2. Students will take the pretest and posttest batteries before and after their game-playing activity, respectively.
2) *Portal 2*—180 students will play Portal 2 for eight hours, spanning 2-3 weeks. Stealth assessments will record performance indicator data, and update competency estimates in real time. We anticipate about 30 minutes for learning controls and warm up, during which time the stealth assessment data will *not* be used for assessment purposes. Students will take the pretest and posttest batteries before and after Portal 2 gameplay, respectively.

Data Analyses

Reliability and validity of stealth assessments. Split-half reliabilities will be calculated across all evidence indicators in Portal 2 for each competency. Thus if there are, say, 40 evidence indicators of persistence in Portal 2, we will randomly split those indicators into two indicator sets (compiled for all students) and correlate the competency estimates between the two sets. We expect to see strong correlations between the indicator sets for each competency ($r \geq .80$), indicating good reliabilities. We will evaluate construct validity of the stealth assessments by computing correlations between the competency estimates from the game (for each of our three competencies) and the associated traditional measures. We expect to see moderate positive correlations between the stealth assessment estimates and the traditional assessment scores ($r = .40$ - $.60$). We also expect to see relatively small relations between our three stealth assessment scores.

The predictive validity of the stealth assessments will be evaluated by computing hierarchical regression analyses to investigate how well the stealth assessment estimates predict student GPA and FCAT scores over and above the predictive ability of the traditional measures. We expect that the stealth assessments will predict GPA and FCAT scores 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 assessment estimates compared to the three traditional assessment scores.

Learning gains. To investigate learning with our external measures, we will compute a 2 × 2 mixed model ANOVA to compare pretest-posttest learning (within-subjects measure) in Portal 2 versus the control group (between-subjects measure). We expect to see a significant difference between pretest and posttest for certain competencies of the Portal 2 group relative to the control group differences. Specifically, we posit that problem solving skills will be significantly enhanced from playing Portal 2, and possibly spatial abilities and persistence will improve as well. Covariate measures (e.g., gender, grade) will be used to control for any individual differences variance.

To examine learning within our stealth assessments in the game, we will test a variety of methodologies using both Bayes net probability estimates and mean score estimates across a set of observables identified in gameplay in Portal 2. These estimates will be calculated at different time points in the 8-hour gameplay to evaluate growth over time. For example, to examine learning with our internal (stealth assessment) measures using Bayes nets, we will need to first reduce the probability estimates per competency (e.g., high, medium, and low levels) to a single number. To do this we will assign numeric values +1, 0 and -1 to the three states, and compute the expected value. This Expected A Posteriori (EAP) value can also be written as, $P(\theta_{ij} = \text{High}) - P(\theta_{ij} = \text{Low})$, where θ_{ij} 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. We will then be able to calculate growth over time from the EAP values for all three competencies—problem solving, spatial abilities, and persistence.

Discussion and Implications

This research aims to contribute to two areas. First we will be able to provide much-needed empirical findings regarding the learning of three important competencies from a well-designed game. Second, we will validate our stealth assessments in a well-designed game.

This proposed research extends our current research efforts (funded by the Gates Foundation) examining stealth assessments in Crayon Physics Deluxe in several ways. First, we will evaluate stealth assessments in another game—one that is widely used in the general gaming population. Second, we will model and examine a different set of important competencies in Portal 2 (i.e., problem solving skill and spatial ability). Third, we plan to examine the extent to which our persistence models (developed for the Gates project) can be applied in Portal 2, dealing with scalability issues of our stealth assessments. Fourth, we will be evaluating learning more rigorously in the proposed research by including a control condition. Finally, with the proposed research, we will be working directly with a game company that is keenly interested in examining learning from games.

This research has implications for game developers by providing a research foundation and methodology to build educationally-focused video games. For instance, evidence that Portal 2 can be used as the basis for valid assessments and to improve learning would be encouraging news for game developers who want to create games for educational purposes. Additionally, this research may motivate education researchers to work with game developers to create "competency-focused" games (i.e., games that are created for the sole purpose of assessing and supporting a particular competency). As we have pointed out (Shute, Ventura, & Kim, 2011), much can be learned from both researchers and game developers to make successful educationally-focused video games that can have a positive impact on student success—in school and in life.

Timeline

Tasks/Deliverables	Dates
Develop ECD-based Competency and Evidence models	Jan. – Mar. (2012)
Meet with Valve in Seattle	Mar.
Refine ECD models based on meeting	Apr.
Create problems with Portal 2 “mod” tool	Apr. – May
Create task models and select problems (old and new) for pilot testing	May – Jun.
Conduct pilot study at FSU and/or Valve (*n* = 50)	Jun. – July
Analyze results of pilot study	Aug.
Refine ECD models based on pilot study	Sept.
Select traditional assessments for validity research	Oct.
Meet with Valve to finalize problems and stealth assessments	Dec. – Jan. (2013)
Finalize full set of Portal 2 problems	Feb.
Run validity/learning study at FSU (*n* = 200)	Mar. – Jun.
Analyze results	Jun. – Aug.
Write final report	Sept. – Dec.

Project staff

Dr. Valerie Shute (PI) is a Professor 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 research scientist in the Educational Psychology and Learning Systems department at Florida State University. He is a cognitive scientist with expertise in educational technology and noncognitive assessment. Matthew's primary responsibility will be to oversee the design of competency, evidence, and task models. Dr. Fengfeng Ke (Co-PI) is an educational psychologist in the Educational Psychology and Learning Systems department at Florida State University and an expert on game-based learning and educational game design. Fengfeng’s primary responsibility will be to assist the design and implementation of the ECD models, and oversee the learning study for both the control and Portal 2 conditions.

Communications Plan

If our evidence-based stealth assessment methodology is found to be valid and reliable, we plan to make the process and the models broadly available so that the work will continue in other research and real-world settings. 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 findings and models via the new "games, learning, and assessment" research area that has emerged from the recent Gates-MacArthur Workshop on the topic. Finally, we plan to publish our findings in peer-reviewed journals and make our models available to other researchers via this more traditional, scholarly venue.

References

Abe, J. A. A. (2005). The predictive validity of the Five-Factor Model of personality with preschool age children: A nine year follow-up study. *Journal of Research in Personality, 39* (4), 423-442.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, A. & Bavelier, D. (2011). Action game play as a tool to enhance perception, attention and cognition. In S. Tobias & D. Fletcher (Eds), *Computer games and instruction* (pp. 307-330). Information Age Publishing: IAP, Charlotte, NC.

Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly, 23*(3), 299-313. doi: 10.1016/j.ecresq.2008.03.001

Bodrova, E., & Leong, D. J. (2007). *Tools of the Mind: The Vygotskian approach to early childhood education* (2nd ed.), Upper Saddle River, NJ: Prentice-Hal.

Bransford, J. D. & Stein, B. S. (1993). *The ideal problem solver* (2nd ed.), New York: Freeman.

Bransford, J. D., Vye, N., Kinzer, C., & Risko, V. (1990). Teaching thinking and content knowledge: Toward an integrated approach. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 381-413). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bransford, J. D., Arbitman-Smith, R., Stein, B. S., & Vye, N. J. (1985). Improving thinking and learning skills: An analysis of three approaches. In J. W. Segal, S. F. Chipman, & R. Glaser (Eds.), *Thinking and learning skills: Relating instruction to research* (Vol. 1, pp. 133-206). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review, 6*, 345-375.

Charsky, D. (2010). Making a connection: Game genres, game characteristics, and teaching structures. In R. Van Eck (Ed.), *Gaming & cognition: Theories and perspectives from the learning sciences* (pp. 189 – 212). Hershey, PA: IGI Global.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience.* New York: Harper and Row.

Darken, R., Allard, T., & Achille, L. (1998). Spatial orientation and wayfinding in large-scale virtual spaces: An introduction. *Presence 7*(2), 101-107.

De Fruyt, F. & Mervielde, I. (1996). Personality and interests as predictors of educational streaming and achievement. *European Journal of Personality, 10*, 405-425.

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, *318*(5855), 1387-1388.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92(*6), 1087-1101.

Duncker, K. (1945). On problem-solving. *Psychological Monographs, 58*, 5, 1-113.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 685-706). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory: Toward a science of expert and exceptional performance. *Current Directions in Psychological Science, 16*(6), 346-350.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Gee, J. P. (2005). What would a state of the art instructional video game look like? *Innovate, 1*(6). Retrieved June 20, 2011, http://www.innovateonline.info/index.php?view=article&id=80

Glaser, R., Pellegrino, J. W., & Lesgold, A. M. (1978). Some directions for a cognitive psychology of instruction. In A. M. Lesgold, J. W. Pellegrino, S. O. Fokkema, & R. Glaser (Eds.), *Cognitive psychology and instruction* (pp. 495-517). New York: Plenum.

Hayes, J. R. (1989). *The complete problem solver* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates.

Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T. and Lovelace, K. (2006) Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence, 34*, 151-176.

Hegarty, M., Crookes, R. D., Dara-Abrams, D. & Shipley, T. F., (in press). Do all science disciplines rely on spatial abilities? Preliminary evidence from self-report questionnaires. To appear in C. Hölscher, T. F. Shipley, M. Olivetti, J. Bateman, & N. Newcombe (Eds.), *Spatial Cognition VII***.** Lecture Notes in Computer Science. Springer.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*, 175-191.

Jonassen, D. H. (2002). Integration of problem solving into instructional design. In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design & technology* (pp.107-120). Upper Saddle River, NJ: Merrill Prentice Hall.

Jonassen, D. H., Marra, R. M., and Palmer, B. (2004). Epistemological development: An implicit entailment of constructivist learning environments. In N. M. Seel & S. Dikjstra (Eds.) *Curriculum, plans and processes of instructional design: International perspectives* (pp. 75-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Kozhevnikov, M. (2008). U.S. Patent: *Three-dimensional perspective taking ability assessment tool.* U.S. Patent #2010/0075284. http://www.freepatentsonline.com/20100075284.pdf

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009) Empirical identification of the major facets of Conscientiousness. *Learning and Individual Differences, 19*, 451-458.

Maguire, E. A., Frackowiak, R. S. J., & Frith, C. D. (1997). Recalling routes around London: Activation of the right hippocampus in taxi drivers, *The Journal of Neuroscience, 17*(18), 7103-7110.

Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias & D. Fletcher (Eds), *Computer games and instruction* (pp. 281-306). Information Age Publishing: IAP, Charlotte, NC.

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

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.) *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

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

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

Pellegrino, J. W. (1984, April). *Information processing and intellectual ability.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification.* Washington, DC: American Psychological Association.

Polya, G. (1945). *How to solve it.* Princeton, NJ: Princeton University Press.

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

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1-66.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Shiner, R. L., Masten, A. S., & Roberts, J. M. (2003). Childhood personality foreshadows adult personality and life outcomes two decades later. *Journal of Personality, 71*(6), 1145-1170. doi: 10.1111/1467-6494.7106010

Shute, V. J. (1984). *Characteristics of cognitive cartography*. Unpublished Ph.D. Dissertation, Graduate School of Education, University of California, Santa Barbara.

Shute, V. J., Gawlick, L. A., & Gluck, K. A. (1998). Effects of practice and learner control on short- and long-term gain and efficiency. *Human Factors, 40*(2), 296-310.

Shute, V. J., Ventura, M., & Kim, Y. J. (2011). *Synthesis report on the games, learning, and assessment (GLA) Workshop.* Paper prepared for the Gates and MacArthur Foundations.

Sternberg, R. J. (1999). *Handbook of creativity.* Cambridge, UK: Cambridge University Press.

Ruscio, A., M., Amabile, T. M., (1999) Effects of instructional style on problem-solving creativity**,** *Creativity Research Journal, 12,* 251-266.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs 2* (Whole No. 8).

Underdahl, J., Palacio-Cayetano, J., & Stevens, R. (2001). Practice makes perfect: Assessing and enhancing knowledge and problem-solving skills with IMMEX software. *Learning & Leading with Technology, 28*(7), 26-31.

Van Eck, R. (2009). A guide to integrating COTS games into your classroom. In R. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (pp. 179–199). Hershey, PA: Idea Group.

Voss, J. F. (1989). Problem solving and the educational process. In A. Lesgold & R. Glaser (Eds.), *Foundations for a psychology of education* (pp. 251-294). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky*. New York: Plenum.

Aug. 12, 2011

Dear Val,

I am pleased to report that Valve will support the **Stealth Assessment in Portal 2** research proposal to the MacArthur Foundation if awarded per your application.

No monies will be received by Valve, nor will we be funding it beyond in-kind support described in our phone call of this morning to include: cooperation regarding the mechanics of the project, pre-testing on site at Valve and software licenses.

We are delighted to be collaborating with you and the MacArthur Foundation in assessing and promoting crucial life-long skills like problem solving, persistence and spatial abilities.

Best, Leslie Redd

■ Leslie Redd ■ Director of Educational Programs ■ Valve Corporation ■ leslier@valvesoftware.com ■ 425.889.9642 x224 ■

FLORIDA STATE UNIVERSITY SCHOOLS, INC.

3000 School House Road
Tallahassee, FL 32311
(850) 245-3700 FAX (850) 245-3737
www.fsus.fsu.edu

To: Whom It May Concern

From: Dr. Lynn Wicker, Director, Florida State University Schools

Date: August 22, 2011

Re: Approval Notification for Stealth Assessment in Portal 2

This letter serves as approval for the upcoming MacArthur research project entitled *Stealth Assessment in Portal 2* at the Florida State University School (FSUS). We look forward to working with Dr. Valerie Shute (PI) and other project investigators as they design, develop, and evaluate an evidence-based, diagnostic stealth assessment system embedded into the fabric of a gaming environment. Additionally, the project will include the following components:

- A heterogeneous sample of high school students (i.e., grades 10-12; 15-19 years old). All studies will be conducted in the Stone Building at FSU (after school and weekends).
- Students will be randomly assigned to play either Portal 2 or another video game across several 1-2 hour sessions. The students will complete a set of assessments before and after the gaming activities—measuring problem solving, spatial ability, and persistence. Students will also complete a short questionnaire about their prior gaming experience.
- FSUS will provide state test score data (FCAT) and GPA to the researchers.
- The PI will be responsible for ensuring that all of FSU's IRB policies and procedures are followed.

We look forward to participating in the success of this innovative project.

Lynn A Wicker

Lynn A. Wicker, Director