Lessons Learned and Best Practices of Stealth Assessment

Lubin Wang, Florida State University, Tallahassee, FL, USA

Valerie Shute, Florida State University, Tallahassee, FL, USA

Gregory R. Moore, Florida State University, Tallahassee, FL, USA

ABSTRACT

Stealth assessment provides an innovative way to assess and ultimately support knowledge, skills, and other personal attributes within learning and gaming environments without disrupting students' flow. In this paper, the authors briefly discuss two challenges they encountered during the development of stealth assessments in two past projects (i.e., utility issues related to log files and validation issues related to in-game measures). They also present successful examples of designing and testing stealth assessments and describe the steps they are taking to apply the lessons they have learned to the ongoing development of a stealth assessment for problem solving skills. The authors conclude with suggestions for future research.

Keywords: *21st Century Skills, Bayesian Networks, Evidence-Centered Design, Game-Based Learning, Log File Analysis, Stealth Assessment*

1. INTRODUCTION TO EVIDENCE-CENTERED DESIGN AND STEALTH ASSESSMENT

Today's students are expected to develop 21st century skills, such as problem solving, creativity, and critical thinking (Partnership for 21st Century Learning, 2012). Such higher-order skills are necessary to be successful and productive in school, work, and life in general. It is thus important for educators to be able to accurately assess students on these complex skills. Assessments can help educators determine not only students' current levels of these competencies, but also their strengths and weaknesses on particular facets of the skills. This information can assist educators in supporting their students to develop 21st century skills, as well as other important competencies such as content knowledge and dispositions. However, traditional formats for assessing learning and achievement, such as multiple-choice tests, often measure superficial skills and are stripped of the context in which knowledge and skills are applied (Shute, Leighton, Jang, & Chu, in press). Thus, an ongoing problem in education involves finding more authentic and valid, yet efficient, ways to assess students on these complex competencies. Stealth assessment

DOI: 10.4018/IJGCMS.2015100104

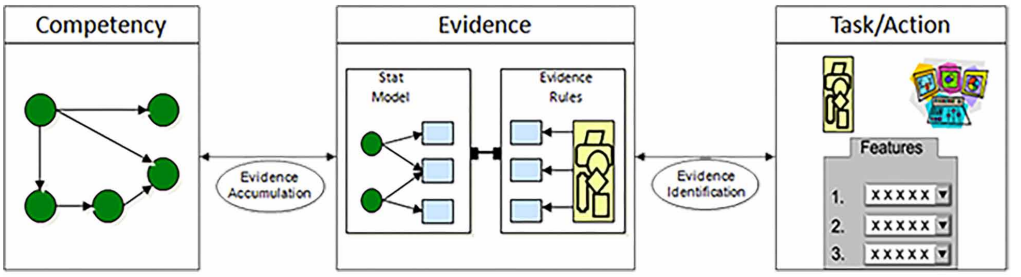
(Shute, 2011) has been proposed as one of the most promising methods for assessing complex skills. It is the process of embedding assessments seamlessly into a computer-based learning or gaming environment such that the learner is unaware he or she being assessed.

Researchers generally agree that the development of an assessment has to follow a principled assessment design framework (AERA, APA, NCME, 1999; Kane, 2006) to be valid and reliable. Some leading principled assessment design frameworks include evidence-centered design (ECD), cognitive design system (CDS), and assessment engineering (AE). These three design frameworks are similar in their end goals, but vary in the processes they use to arrive at the goals (Shute, Leighton, Jang, & Chu, in press). In this paper, we discuss the hurdles we faced when using the evidence-centered design framework to implement stealth assessment and how we overcame those hurdles. Based on these hurdles, we make recommendations for stealth assessment best practices. We also present an ongoing project in which we are applying the lessons we have learned to more effectively and efficiently develop and implement stealth assessment.

Evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003) is a framework that can be used to design valid assessments for measuring students' knowledge, skills, and other attributes. The framework is made up of three main models that work together: the competency model, the evidence model, and the task model (see Figure 1). The competency model contains the variables that characterize the competency of interest. Beliefs about students' status on the variables are represented by probability distributions that can be updated whenever new information is acquired. When a competency model is instantiated with data relating to a student's performance, it is called the student model. The task model specifies features of the tasks that students will undertake to provide evidence about target competencies. The features include the materials to be presented to students and the work products expected from them. The competency model and the task model are both connected to the evidence model, which provides a statistical link between the two. The evidence model consists of (a) evidence rules that convert the work products to observable variables, and (b) the statistical model that defines the statistical relationships between the observable variables and competency variables. In this way, evidence about the observable variables will update the competency model so that it accurately reflects the student's knowledge, skills, or other attributes at any time and at a fine grain size. The focus on the evidentiary link between the claims made about an examinee's competency and the collected evidence is the main feature of ECD that distinguishes it from other leading principled assessment design frameworks. Thus, creating an assessment using ECD allows one to evaluate the performance data that results from engaging in various tasks and, in turn, make inferences about various competencies (e.g., problem solving skills). Furthermore, coupling ECD with technology-enhanced environments (e.g., games), allows one to collect copious amounts of data and make valid inferences relative to the competencies.

Over the past decade, we have been using games as our preferred vehicle for assessing higher-order competencies. There are two main reasons for this choice. First, video games are becoming increasingly popular, especially among teenagers (Lenhart et al., 2008). The vast majority of teenagers play games in their free time, as they find the activity engaging and enjoyable. The meaningful contexts provided by games allow the embedded assessment engine to obtain solid, cohesive, and detailed information about players' competencies. Moreover, players may not be aware of the fact that they are being assessed, which frees them from the anxiety commonly associated with traditional tests. Second, Gee (2003) and other scholars have suggested that games can help students develop problem-solving skills, as well as other valuable 21st century competencies. In a well-designed game, players need to apply these competencies to complete the goals of the game. At the same time, games provide immediate feedback in the form of scores

Figure 1. The three main models of ECD (from Mislevy, Steinberg, & Almond, 2003)



or the progress of the player's in-game character. Therefore, well-designed games allow players to both learn valuable competencies and be assessed at the same time.

To assess target competencies unobtrusively in games, we have used stealth assessment, which is a specialized implementation of ECD. Stealth assessment helps to reduce test anxiety and maintain learners' engagement. It works as follows. As students interact with tasks/problems in a game during the solution process, they are providing a continuous stream of data, which is captured in a log file and then analyzed by the evidence model. The results of this analysis are data (e.g., scores) that are passed to the competency model, which statistically updates the claims about relevant competencies in the student model. The estimates of competency levels can then be used diagnostically and formatively to provide feedback and other forms of learning support to students as they continue to engage in gameplay. This process, of making valid assessments and then using that information as the basis for offering learning support to the student, is important in supporting the growth of competencies.

2. LESSONS LEARNED AND BEST PRACTICES

To date, we have developed a number of stealth assessments for use in different games to examine various competencies. For example, we developed stealth assessments to measure problem solving and spatial skills in *Portal 2* (Shute, Ventura, & Ke, 2015; Shute & Wang, in press), causal reasoning in the *World of Goo* (Shute & Kim, 2011), and systems thinking in *Taiga Park* (Shute, Masduki, & Donmez, 2010). Additionally, we designed three stealth assessments to measure various cognitive and noncognitive variables in a game called *Physics Playground* (formerly called Newton's Playground; see Shute & Ventura, 2013). The focal competencies included persistence (Ventura, Shute, & Small, 2014; Ventura, Shute, & Zhao, 2012), qualitative physics knowledge (Shute, Ventura, & Kim, 2013), and creativity (Kim & Shute, in press). From these design and development efforts, we have learned a number of useful lessons about developing and applying stealth assessments.

In this section, we share some lessons learned about stealth assessment that come from our work on two past research projects. We also make recommendations based on our experiences and present the progress of a current research project applying these lessons learned. The first project we examine used stealth assessment to examine problem-solving skill, spatial skill, and persistence in the popular commercial game, *Portal 2* (developed by Valve Corporation). The second project on the other hand, used stealth assessment to examine physics understanding, creativity, and persistence relative to validity, learning, and enjoyment in the game *Physics Playground*. The project we present at the end of this section is an ongoing joint effort between our

research team and GlassLab (see https://www.glasslabgames.org/). We are developing a stealth assessment of problem solving skills and embedding it directly into the popular game, *Plants vs. Zombies 2* (developed by Popcap Games and Electronic Arts). We start with our lessons learned and the recommendations that stem from those lessons.

2.1. Game Logging Systems

2.1.1. Lesson: Make Sure that the Log Files are Manageable and/or Customizable

One lesson related to employing stealth assessment concerns game logging systems, which play a key role in the first phase of the assessment cycle. In stealth assessment, the role of a game logging system is to record performance data as players advance in a game. The logs are then analyzed and key information is extracted to inform a player's target competency (or competencies). For the Portal 2 project, we used the commercial game without any modification. Portal 2 has a built-in logging system, so when players engage in gameplay, their in-game behaviors are recorded by log files in real-time. Based on the stealth assessment cycle, our initial plan was to (a) extract evidence of the three competencies from the gameplay log files (via indicators, like the coordinates of portal shots per level), (b) score the evidence based on predetermined scoring rules, (c) accumulate scores in Bayesian networks, or Bayes nets, and (d) update the estimates of students' competencies expressed as probability distributions in the competency model (We talk about how these steps are usually performed in the PvZ2 example at the end of this section). Unfortunately, we encountered difficulties at step (a) (i.e., extracting meaningful information from the log files). The code was developed and compiled by the development team at Valve and was not intended for outside use or for assessment purposes. Consequently, although we could access each student's log file data (via the developer's console), we were not able to obtain a complete coding scheme. In addition, the logging system recorded every single action and event in the game in milliseconds. As a result, the stream of data being logged became unmanageable after only a short period of gameplay. Figure 2 displays a screen capture of a tiny part of the log file. The snapshot shows the code produced by the logging system at 111.80 seconds, which includes around 50 lines of code. In the end, we managed to extract a set of actions (e.g., average number of portals shot, average time per level) from each player's log file. However, we did not have enough time to extract as much evidence as we intended.

This experience has implications for the future selection of games for assessment purposes. Many researchers may be tempted to use the readily available and appealing commercial games to avoid the hassle of creating a new game from scratch. However, if a researcher wants to create stealth assessments within an existing commercial game, she or he must first make sure that the coding in the log files is simple enough to understand or that the coding scheme is available from the game developer so that changes can be made to the information that is being captured. At the same time, this experience revealed one of the advantages of homemade games. Not only can researchers design the content and presentation of the games the way they want, but they are also able to customize the format of the log files at the outset of game and assessment design.

2.1.2. Best Practice: Include Well-Organized, Necessary Data in the Log File

In the Physics Playground project, we designed the game such that it would automatically upload session logs to a server. A session is defined as the actions a player takes between login and logout. The log files were designed to be simple enough to retrieve useful information quickly. Additionally, we ensured that all of the data we needed were captured in the log files. Figure 3

Figure 2. A snapshot of a Portal 2 log file at one point in time

```
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) ->
(!self,CallScriptFunction) (CatcherPowerOff)
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) -> (lasercatch54-proxy,OnProxyRelay2) ()
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) -> (lasercatch54-proxy,OnProxyRelay4) ()
(111.83) input indicator_toggle99-indicator_on_rl: indicator_toggle99-indicator_on_rl.EnableRefire()
(111.83) input pistonlift3-relay_lift_down: pistonlift3-relay_lift_down.EnableRefire()
(111.83) input pistonlift3-branch_toggle: pistonlift3-cube_enable_motion_trigger.Disable()
(111.83) input lasercatch54-laser_catcher: lasercatch54-laser_catcher.CallScriptFunction(CatcherPowerOff)
(111.83) input lasercatch54-laser_catcher: lasercatch54-proxy.OnProxyRelay2()
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) -> (pistonlift51-proxy,OnProxyRelay2) ()
(111.83) input lasercatch54-laser_catcher: lasercatch54-proxy.OnProxyRelay4()
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) -> (indicator_toggle100-
proxy,OnProxyRelay2) ()
unhandled input: (StopSound) -> (lasercatch54-laser_catcher_music_*), from
(prop_laser_catcher,lasercatch54-laser_catcher); target entity not found
unhandled input: (StopSound) -> (lasercatch10-laser_catcher_music_*), from
(prop_laser_catcher,lasercatch54-laser_catcher); target entity not found
unhandled input: (PlaySound) -> (lasercatch10-laser_catcher_music_1), from
(prop_laser_catcher,lasercatch54-laser_catcher); target entity not found
(111.83) input lasercatch54-laser_catcher: pistonlift51-proxy.OnProxyRelay2()
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) -> (pistonlift51-counter,Subtract) (1)
(111.83) input lasercatch54-laser_catcher: indicator_toggle100-proxy.OnProxyRelay2()
(111.83) output: (prop_laser_catcher,Lasercatch54-laser_catcher) -> (indicator_toggle100-
indicator_off_rl,Trigger) ()
(111.83) input lasercatch54-laser_catcher: pistonlift51-counter.Subtract(1)
(111.83) output: (math_counter,pistonlift51-counter) -> (pistonlift51-branch_toggle,ToggleTest) ()
(111.83) input lasercatch54-laser_catcher: indicator_toggle100-indicator_off_rl.Trigger()
(111.83) output: (logic_relay,indicator_toggle100-indicator_off_rl) -> (indicator_toggle100-
texture_toggle,SetTextureIndex) (0)
(111.83) input pistonlift51-counter: pistonlift51-branch_toggle.ToggleTest()
(111.83) output: (logic_branch,pistonlift51-branch_toggle) -> (pistonlift51-lift_fizzler,Disable) ()
(111.83) output: (logic_branch,pistonlift51-branch_toggle) -> (pistonlift51-
cube_enable_motion_trigger,Disable,0.0) ()
(111.83) output: (logic_branch,pistonlift51-branch_toggle) -> (pistonlift51-
cube_enable_motion_trigger,Enable) ()
(111.83) output: (logic_branch,pistonlift51-branch_toggle) -> (pistonlift51-relay_lift_up,Trigger) ()
(111.83) input indicator_toggle100-indicator_off_rl: indicator_toggle100-
texture_toggle,SetTextureIndex(0)
(111.83) input pistonlift51-branch_toggle: pistonlift51-lift_fizzler.Disable()
(111.83) input pistonlift51-branch_toggle: pistonlift51-lift_fizzler.Disable()
(111.83) input pistonlift51-branch_toggle: pistonlift51-cube_enable_motion_trigger.Enable()
(111.83) input pistonlift51-branch_toggle: pistonlift51-relay_lift_up.Trigger()
(111.83) output: (logic_relay,pistonlift51-relay_lift_up) -> (pistonlift51-
counter_lift_target_up,GetValue) ()
(111.83) output: (logic_relay,pistonlift51-relay_lift_up) -> (pistonlift51-
counter_lift_level_up,GetValue) ()
(111.83) input pistonlift51-relay_lift_up: pistonlift51-counter_lift_target_up.GetValue()
(111.83) output: (logic_relay,pistonlift51-relay_lift_up) -> (pistonlift51-case_lift_target_up,InValue) ()
(111.83) input pistonlift51-relay_lift_up: pistonlift51-counter_lift_level_up.GetValue()
(111.83) output: (logic_relay,pistonlift51-relay_lift_up) -> (pistonlift51-case_lift_level_up,InValue) ()
(111.83) input pistonlift51-relay_lift_up: pistonlift51-case_lift_target_up.InValue(2)
(111.83) output: (logic_case,pistonlift51-case_lift_target_up) -> (pistonlift51-
lift_platform,SetPosition) (0.5)
(111.83) input pistonlift51-relay_lift_up: pistonlift51-case_lift_level_up.InValue(1)
(111.83) output: (logic_case,pistonlift51-case_lift_level_up) -> (pistonlift51-
branch_segment_1_up,Test) ()
(111.83) input pistonlift51-case_lift_target_up: pistonlift51-lift_platform.SetPosition(0.5)
(111.83) input pistonlift51-case_lift_level_up: pistonlift51-branch_segment_1_up.Test()
(111.83) output: (logic_branch,pistonlift51-branch_segment_1_up) -> (pistonlift51-
counter_lift_level_down,SetValueNoFire) (1)
(111.83) output: (logic_branch,pistonlift51-branch_segment_1_up) -> (pistonlift51-lift_segment_1,Open) ()
(111.83) input pistonlift51-branch_segment_1_up: pistonlift51-counter_lift_level_down.SetValueNoFire(1)
(111.83) input pistonlift51-branch_segment_1_up: pistonlift51-lift_segment_1.Open()
```

displays a sample log file from a single level in Physics Playground. It logged events such as the entrance to a particular playground and level, the start time, the time spent interacting with the level, the number of objects created, the number of restarts, the agents used, whether the player solved the level or not, and if so, whether she received a gold or silver trophy (see Shute & Ventura, 2013 for details).

2.2. Choosing External Measures to Validate Stealth Assessments and Test Transfer

The first thing to do after developing a stealth assessment is to test for construct (or convergent) validity. That is, we need to ensure that the stealth assessment actually measures what it

Figure 3. A snapshot of a Physics Playground level log file


```
},
"event_1" : {
        "type" : 3,
        "type_string" : "Enter Room",
        "time_stamp" : 93.586998,
        "room_name" : "Playground 4"
},
"event_2" : {
        "type" : 2,
        "type_string" : "Play Level",
        "time_stamp" : 125.526001,
        "log_file_name" : "ALLEN_314_6_play1.replay",
        "level_path" : ".\\levels\\p3\\double bounce.level",
        "game_time" : 69.167999,
        "pause_time" : 0,
        "restart_count" : 0,
        "object_count" : 2,
        "object_limit_count" : 0,
        "nudge_count" : 0,
        "erase_count" : 2,
        "pin_count" : 1,
        "agent_vector" : "",
        "ball_trajectory" : "<0.694, -0.293> <0.653, -0.186> ...
        "silver" : true,
        "gold" : false,
        "solved" : true
},
```


is supposed to be measuring. External measures that can be used for such validation include well-established standardized tests of the focal construct, other relevant assessments that have been validated and have reasonable reliabilities, and self-report surveys/questionnaires related to the target competency.

In general, we would not recommend using self-report measures as an external test in a validation study. One problem with self-report measures is that they suffer from what is often called "social desirability effects" (Paulhaus, 1991). This refers to the tendency for people to answer in line with what society or the researchers view as favorable rather than their actual beliefs. This effect can lead to the inflation of scores related to good behaviors and/or the reduction of scores related to bad behaviors in the self-report. Another issue with self-report is that people sometimes have different conceptual understandings of the questions (e.g., what it means to "work hard" as part of a persistence question), leading to low reliability and validity (Lanyon & Goodstein, 1997). Finally, self-report items often require that individuals have explicit knowledge of their skills and dispositions (see, e.g., Schmitt 1994), which is not always the case. People may find it difficult to accurately score themselves along the scales provided in a self-report (e.g., the ambiguity between good and excellent) because they possess different levels of knowledge about themselves and/or different personalities (e.g., some are more humble while others are more confident about themselves). All of these weaknesses may undermine self-report as an ideal external measure.

Another difficulty typically associated with the selection of external measure(s) is the detection of transfer beyond the game environment (e.g., Boot, Kramer, Simons, Fabiani, & Gratton, 2008). This difficulty is likely caused by choosing the wrong type of external assessment for the transfer task. That is, traditional types of assessment usually consist of multiple-choice questions, true or false, short answers, or self-report surveys without context. Many people experience test

anxiety with these tests, which may influence one's performance. Also, the scope of traditional test items may not be sufficient to cover all that is taught by the treatment because of the limited number of test items that can be presented. Occasionally, the dimensionality of external measures may not be a good fit to the internal measures because many complex competencies (e.g., creativity and problem solving skills) are very broad and include many facets. Researchers must select external measures with caution to make sure that the external measures align with the in-game (or stealth) measures. One of the external measures we employed in the Portal project suffered from this misalignment issue.

2.2.1. Lesson: Misalignment of External Measures with in-Game Measures

In 2014, 77 undergraduate students from various majors at a university located in the southeastern U.S. participated in our Portal 2 study. Participants were randomly assigned to the experimental group, playing Portal 2 (42 students), or the control group, playing Lumosity (35 students). Lumosity is an online commercial training program that claims to support the development of various cognitive skills, such as problem solving, flexibility, attention, and information processing speed. Participants played their assigned game for 8 hours across four sessions in our laboratory. Before playing the game, participants completed an online set of problem solving and spatial ability pretests. Then, during the last session, subjects completed a set of matched posttests covering the same skills.

The 64 levels in Portal 2 provided players with extensive practice solving complex problems and engaging in spatial navigation. The game environment was dynamic and required players to generate new knowledge as they advanced through the game. Later levels could only be solved with previously acquired knowledge and skills. Frequently, the game required players to use a tool in a new way, different from how it was learned or used previously. Our in-game measures of problem solving included variables such as the (a) total number of levels solved (more is better), (b) average number of portals shot (less is better), and (c) average time spent solving each level (less is better). We selected three external measures of problem solving to validate our in-game measures and to examine learning transfer from playing the games: Raven's Standard Progressive Matrices (Raven, 1941), insight problems (Weisberg & Alba, 1981), and the remote association test (Mednick, 1962).

- **Raven's Progressive Matrices:** Tested each participant's ability to figure out the missing piece of a matrix based on the given pattern(s). We selected 12 items from the Raven's Progressive Matrices test for the pretest and 12 matched items for the posttest. We matched the items in the two tests by difficulty level (as presented in the RPM test kit), choosing 4 easy, 4 medium, and 4 difficult items per form.
- **Insight Problems:** Are similar to riddles in nature. They yield an "Aha" moment once the solution is found (Chu & MacGregor, 2011). Insight problems usually require problem solvers to shift their perspective and think about the obscure features of the given information. For example: *You need to get from one side of a 100-foot wide and 100-foot deep canyon to the other side. All you have is a 12-foot ladder and an endless supply of rope. How will you cross the canyon?* The correct answer is to use the endless supply of rope to fill in the canyon and then walk over to the other side. Such problems require participants to break from routine thinking. We selected 3 insight problems for the pretest and 3 matched ones for the posttest.
- **The Remote Association Test:** Was originally developed to test creative thinking without any need for prior knowledge. Participants are required to come up with the solution word

that can be associated with each of the three provided words in the form of synonymy, a compound word, or semantic association (Chermahini, Hickendorff, & Hommel, 2012). For instance, the word that can be associated with the triad dream/break/light is "day." We selected 5 items for the pretest and 5 matched items for the posttest.

Unfortunately, our selection of external measures of problem-solving skill suffered from one of the circumstances described above. That is, the dimensionality of some of the external tests did not align well with our in-game measures. Data analysis showed that the correlation between overall Portal 2 performance and Raven's Progressive Matrices scores was not significant (r = .02). Portal 2 performance was also not correlated with the remote association test scores (r =.18). However, the correlation between Portal 2 performance and insight problems was significant (r =.38, $p < .05$). We expected that participants in the Portal 2 condition would perform well on the insight problems because the game required players to think outside of the box. One aspect of the Raven's test is that it only examines subjects' ability to reason based on what is provided directly in the problem. It does not test subjects' ability to apply information in a dynamic environment, as is required by Portal 2. The problem with the remote association test is that it placed a high demand on subjects' English language skills, which confounded the results.

2.2.2. Best practice 1: Choose external measures that align with the stealth assessment

In the Portal 2 project described above, we also investigated whether Portal 2 is an appropriate context for assessing and possibly supporting spatial skills. Researchers generally believe that spatial ability is a significant predictor of performance in science, technology, engineering, and mathematics disciplines (Ventura, Shute, Wright, & Zhao, 2013). We decided to study spatial ability because Portal 2 requires one to move through vast and complex environments during gameplay, explicitly requiring the application of spatial skills to proceed and succeed in the game.

One of the external spatial measures we used for our validation test was the Virtual Spatial Navigation Assessment (VSNA; Ventura, Shute, Wright, & Zhao, 2013). The VSNA was developed in Unity and runs in a web browser. We used it to test participants' environmental (i.e., large-scale) spatial ability. Participants had to locate three colored gems scattered in a virtual 3D environment using a first-person avatar. There were two types of environments presented in the VSNA—(a) the indoor environment (maze-like) with multiple rooms connected by hallways, and (b) the outdoor environment with trees, hills, and bushes (see Shute, Ventura, & Ke, 2015 for more details). Participants needed to complete each environment twice. The first time was the training phase, where participants were expected to familiarize themselves with the environment. The second time was the testing phase, where the sole goal was to collect the 3 gems and return to the home base as quickly as possible. The main measure of environmental spatial ability from the VNSA was the student's time to complete the testing phase. The VSNA automatically recorded the time it took a participant to locate each gem and uploaded that information to a server. Students' performance data in Portal 2 (using a composite measure) was significantly correlated with VSNA performance data (r = .34, $p < .05$). Thus, as expected, the VSNA was well aligned with our stealth assessment in Portal 2. In Portal 2, participants were required to navigate 3D environments that became increasingly difficult as they completed more levels. The VSNA also provided easy and hard environments in which participants could explore, memorize landmarks, and search for target objects.

2.2.3. Best Practice 2: Use Performance-based Assessment over Self-Report for Validation Studies

We recently conducted a study with 154 8th and 9th grade students (72 male, 82 female) at a middle school in the southeastern U.S. (Shute, Ventura, & Kim, 2013). Each student played Physics Playground for 4 hours across a two-week period. We developed stealth assessments to measure the students' qualitative physics understanding, creativity, and persistence. This "best practice" section focuses just on the persistence measure. We were interested in persistence because it is an important personal attribute that predicts academic achievement as well as life outcomes (e.g., Poropat, 2009; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). In Physics Playground, persistence was operationalized in the competency model as the average time spent on unsolved problems and the number of revisits to unsolved problems across all sessions. To validate our stealth assessment measure of persistence, we administered a relevant and widely-used self-report survey from IPIP (i.e., the International Personality Item Pool) and a performance-based measure of persistence (i.e., the PBMP; see Ventura, Shute, & Zhao, 2012). Participants completed both tests through a web browser on a laptop in the school's computer laboratory. For the self-report measure, we used 8 items from the IPIP to assess perceived persistence across different situations. Each item was rated on a 1-5 point Likert Scale (from 1 = strongly disagree to 5 = strongly agree). Sample items included "I have patience when it comes to difficult problems," "I enjoy a good challenge," and "I tend to avoid difficult problems."

For the performance-based assessment, we employed two types of tasks—picture comparisons and anagrams. The purpose of both tasks was to test how long participants would spend on particularly difficult (or impossible) problems. Each of the 7 items (3 easy and 4 difficult) of the picture comparison task presented 2 pictures side by side. The picture on the right had certain features that were missing from the picture on the left. Participants were asked to identify all of the missing pieces between the two pictures. They would click any place on the picture and hit "guess" to see if they correctly identified a difference. Alternatively, they might skip the task at any time to advance to the next item. They had up to 3 minutes per item. For 4 of the 7 items, participants were told that there were 4 differences. However, there really were only 3 perceptible differences. The fourth "difference" was actually only a one-pixel deletion and thus was impossible to detect (see Figure 4 for an example). The time spent searching for the missing pieces was recorded as the score of persistence. Similarly, for the anagrams, four of the seven items were very difficult words (selected on the basis of having very low frequency of usage). An example of a very hard anagram item is shown in Figure 5. Each item had a two-minute limit. The time spent on the impossible anagrams was recorded as the score of persistence.

We administered the persistence self-report at the beginning of the first session and the PBMP at the end of the last session. The results show that, among the 70 low performers in Physics Playground (i.e., those who solved fewer levels), the correlation between the self-report of persistence and the stealth assessment of persistence was not significant (r = -.01). However, the correlation between the PBMP score and the stealth assessment measure of persistence was significant (r = .51, p < .01). Similarly, for the 84 high performers in the game, the correlation between the self-report measure of persistence and the stealth assessment measure was not significant (r = -.06), while the correlation between the PBMP score and the stealth assessment was significant (r = .22, p < .05). We calculated the correlations of high and low performers separately because the same level in the game could be less challenging for high performers than for low performers. Thus, high performers did not need to be as persistent as low performers to solve the level. Because being challenged is one of the conditions for eliciting persistence (Ventura, Shute, & Zhao, 2012), there were fewer opportunities to assess persistence via stealth

Figure 4. An impossible item from the picture comparison task

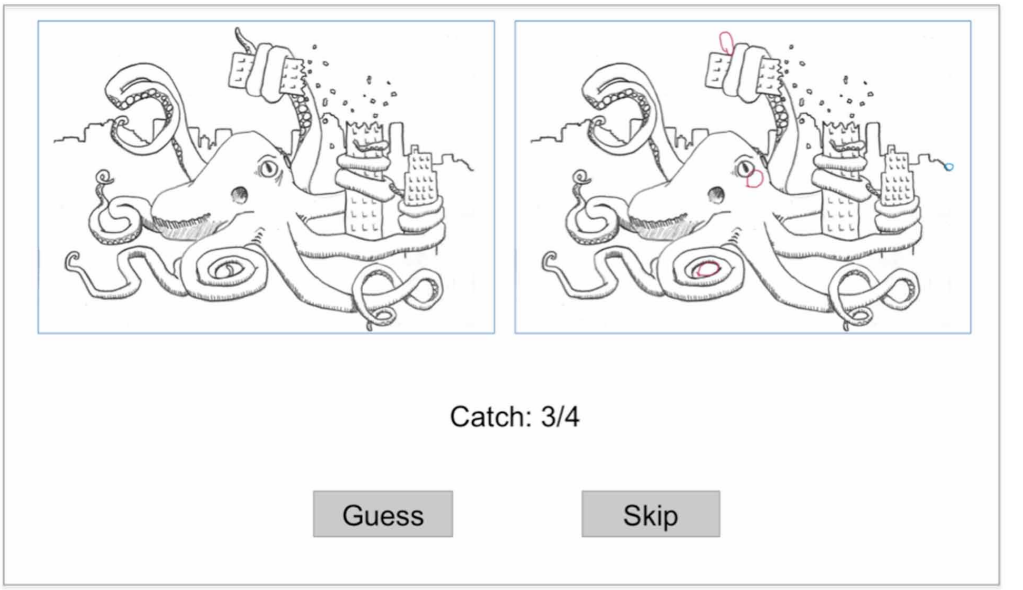
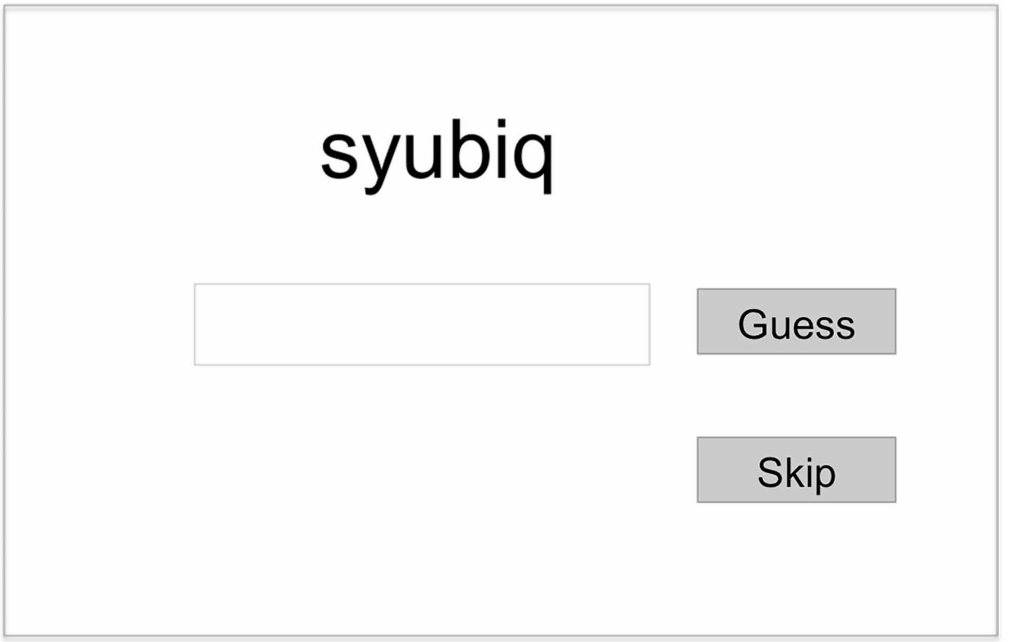


Figure 5. A difficult item from the anagram task



assessment for the high performers than for the low performers. This also likely explains why the correlation between the PBMP score and the stealth assessment for the high performers was lower than that for the low performers.

The PBMP is a good example of an external measure because the format of the test aligned with our stealth assessment. It is performance-based and has a meaningful context in which students are expected to be persistent to solve difficult puzzles. At the same time, since we did not disclose the test's purpose before students took it, the students revealed their true personal attributes rather than changing their responses to what was viewed as desirable.

2.3. Plants vs. Zombies 2 Project

In this section, we present an ongoing project—stealth assessment of problem-solving skills in the game Plants vs. Zombies 2 (PvZ2). We describe how we built different models following ECD and how each model works to generate information we need. As mentioned previously, we are currently collaborating with a team at the GlassLab for this project.

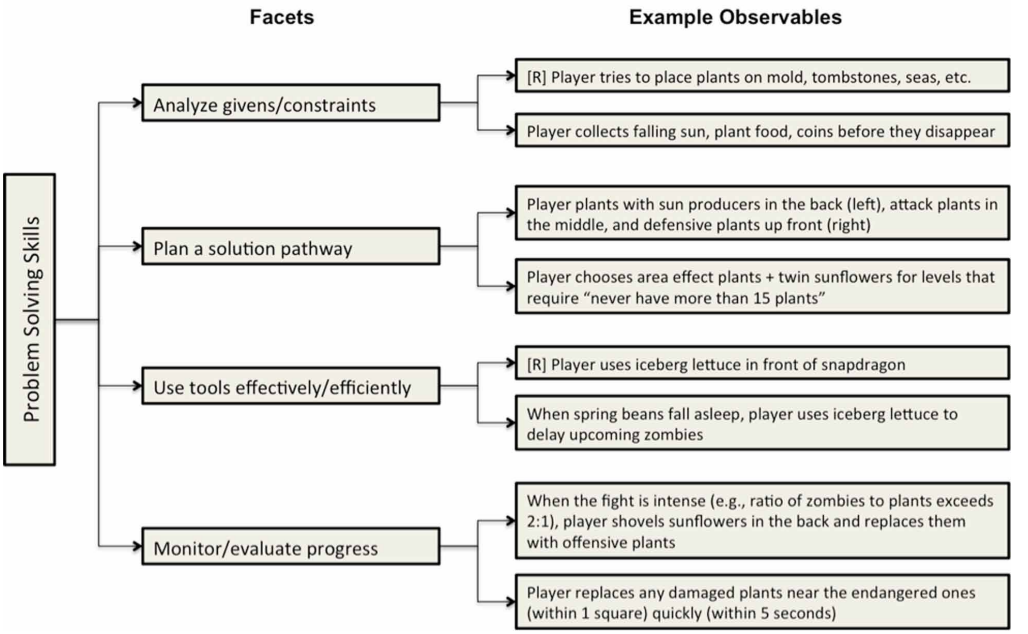
PvZ2 is a tower defense game published by Electronic Arts (EA) that requires players to grow a variety of plants to defeat different invading zombies. GlassLab has obtained the source code of PvZ2 from EA and the technical team at GlassLab is able to customize the log files based on our needs. Before the first step in the stealth assessment cycle, we needed to build models following an assessment design framework to guide our assessment. We selected ECD as the assessment design framework for this project. Because we decided to use an existing game (PvZ2), we did not need to establish a task model as the game and its tasks/levels already existed. Therefore, our focus was on the construction of the competency model and the evidence models.

The first model to build when following the ECD framework is the competency model, which determines the competency that we want to assess in students (problem solving), and the dimensionality of the construct. Towards that end, we reviewed the extensive body of literature on problem solving and came up with four main facets to include in the model: (a) understanding the givens and constraints in a problem, (b) planning a solution pathway, (c) using tools effectively/efficiently during solution attempts, and (d) monitoring and evaluating progress.

After finalizing the competency model, we moved on to the construction of the evidence models. Again, an evidence model consists of (a) evidence rules that convert the work products to observable variables, and (b) the statistical model that defines the statistical relationships between the observable and competency variables. Observable variables provide evidence relative to a student's level on the four facets and overall problem solving skills. After playing through the game and watching solutions to some particularly difficult levels posted on YouTube, we identified a number of in-game indicators that provide evidence for each facet of problem-solving skill (see Figure 6 for an illustration).

Once we determined the observable variables in the game, we needed to decide how to score the observables and establish reasonable statistical relationships between each observable and the associated levels of the competency model variables. We decided that the scoring rule would be based on a tally of relevant instances of observables and then a classification (e.g., into discrete categories such as yes/no, or poor/ok/good/very good). We then constructed Bayesian networks (BNs) to accumulate data and update beliefs in the evidence models. A BN graphically demonstrates the conditional dependencies between different variables in the network. It is composed of both competency model variables (i.e., problem solving and its four facets) and associated observables that are statistically linked to the facets. We constructed a separate BN for each level because the observables change across levels. For instance, a *snapdragon* is a type of

Figure 6. Competency model of problem solving skills and a few example indicators (where [R] refers to reverse-coded indicators)



plant that is locked until the second world. Therefore indicators associated with the snapdragon will not appear in the network until it is unlocked in the game.

Estimates related to players' problem solving skills are updated as ongoing evidence accrues from their interactions with the game. For example, the third facet of problem solving is the ability of a player to use tools effectively and efficiently. One of the plants in the game is iceberg lettuce, which can be used to freeze an incoming zombie temporarily, thus delaying the zombie's attack (see the right side of Figure 7 for the results of zombies coming in contact with iceberg lettuce).

The snapdragon plant mentioned previously breathes fire to burn approaching zombies. Both of these plants (and many others) serve to thwart the onslaught of zombies, and are thus considered valuable resources or tools, if used properly. However, consider the case where a player plants iceberg lettuce in front (i.e., to the right side) of a snapdragon, close to the incoming zombies. That action would indicate poor tool usage because the fire from the snapdragon would melt the ice from the iceberg lettuce immediately, rendering it useless. If a player makes this unfortunate positioning, the log file captures the positioning information and communicates to the evidence model about the ineffective tool use, which in turn updates the estimates about the student's current state of problem-solving skill.

In Table 1, notice the row for indicator #37: *Player plants iceberg lettuce within range of a snapdragon attack (2x3 square space in front of a snapdragon)*. This entry shows how the game log communicates with the node of this indicator in the BN following the evidence rules we set. When a player executes the action of planting an iceberg lettuce in the game, the scripts in the game logging system command a check for a snapdragon in nearby tiles. At the end of a level, the number of iceberg lettuces planted in the range of a snapdragon is divided by the total

Figure 7. Iceberg lettuce in PvZ 2

number of iceberg lettuces planted. Because this is an undesirable action (reversely coded), a *lower* ratio represents better performance. For this indicator, performance is categorized into one of four levels—poor/ok/good/very good. If the ratio falls within [0, 0.25], then this evidence corresponds to the "very good" state in the node in the BN (indicator #37 in Figure 8), given the reverse coding. Similarly, if the ratio falls within [0.26, 0.5], it corresponds to the "good" state of the node; if the ratio falls within [0.51, 0.75], it corresponds to the "ok" state of the node; and if the ratio falls within [0.76, 1], it corresponds to the "poor" state of the node in the network.

The statistical relationships (prior probability distributions) involving indicator #37 and its associated competency variable "efficient/effective tool use" are defined by a conditional probability table (CPT). Table 2 shows the conditional probability table for indicator #37 in level 7 of the Pirate Seas. For example, the value 0.53 in the first cell means that if the player is (theoretically) high on effective/efficient tool use, the likelihood that he or she will rank in the best state "very good" of indicator #37 is 0.53. When evidence about a student's observed results on indicator #37 arrives from the log file, the estimates on his ability to use tools effectively/efficiently will be updated based on Bayes theorem. We configured the distributions of conditional probabilities for each row based on Samejima's graded response model, which includes the item response theory parameters of discrimination and difficulty (see Almond et al., 2001; Almond, 2010; Almond, Mislevy, Steinberg, Williamson, & Yan, 2015).

The discrimination estimate for indicator #37 was set to 0.3 (i.e., low). Discrimination in game-based assessment is expected to be low because of the many confounds involved (Almond, Kim, Shute, & Ventura, 2013). The difficulty for the best state "very good" was set to 0, the difficulty for the second best state "good" was set to -1, and the difficulty for the third state "ok" was set to -2 (i.e., this is a fairly easy item). These parameters were initially determined by a learning scientist, a game expert, and a psychometrician. The CPTs were later calibrated via empirical data collected from a pilot study using the game. The values of the discrimination and difficulty parameters for each indicator in each level were recorded in an augmented Q-matrix

Table 1. The communication between log files and relevant Bayes net nodes (facets)

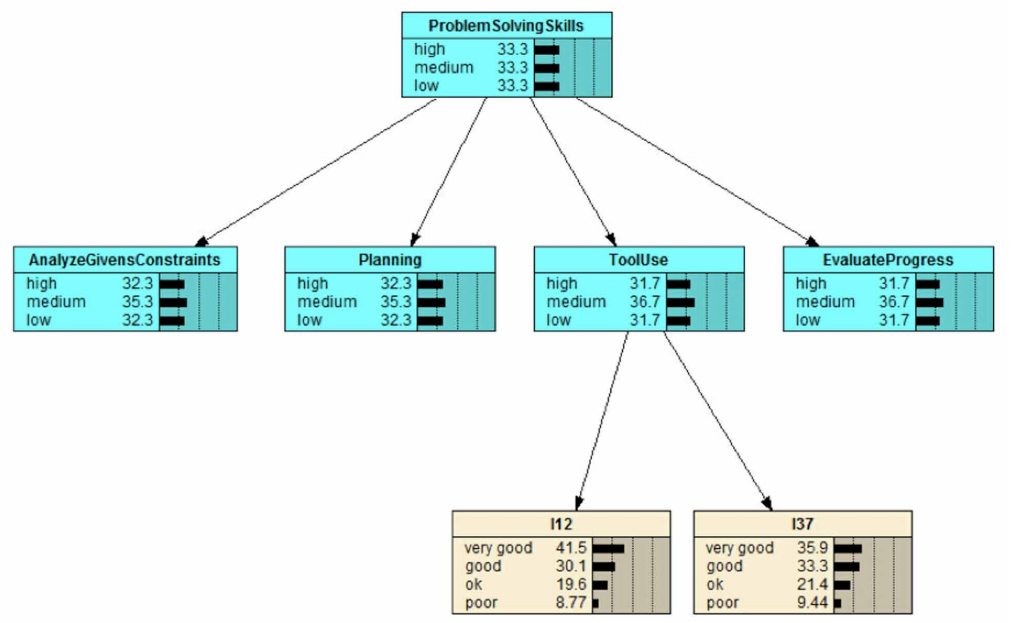
Facets	Indicator #	Indicators	Telemetry event(s) used	Tech Implementation Specifications
Efficient/ effective tool use	37	Player plants iceberg lettuce within range of a snapdragon attack (2x3 square space in front of a snapdragon) [R]	Indicator_planted_iceberg_in_ snapdragon_range	When player plants an iceberg lettuce, check nearby tiles for a snapdragon.
				Ratio = the number of iceberg lettuces planted in the range of a snapdragon/the number of iceberg lettuces planted. Ratio to State: $0 \leq x \leq 0.25$ "very good" $0.26 \leq x \leq 0.50$ "good" $0.51 \leq x \leq 0.75$ "ok" $0.76 \leq x \leq 1.0$ "poor"
	12	Use plant food when there are < 3 zombies on the screen (unless used with sunflowers/ twin sunflowers to get extra sun) [R]	Indicator_percent low_danger_ plant_food_usage.	Ratio = # of plant food used when there are <3 zombies on the screen / total # of plant food used. Ratio to State: $0 \leq x \leq 0.25$ "very good" $0.26 \leq x \leq 0.50$ "good" $0.51 \leq x \leq 0.75$ "ok" $0.76 \leq x \leq 1.0$ "poor"

for possible future adjustment (Almond, 2010). In our Q-matrix, the rows represent the indicators applicable in each level, and the columns represent the four facets of problem solving.

Figure 8 presents a fragment of the problem-solving evidence model, with four main facets and two example indicators of effective tool use (i.e., indicators #37 and #12). We are using the program *Netica* (by Norsys Software Corporation) to construct and compile the network. We selected this software because the user interface is intuitive for drawing the networks. Additionally, the API has been optimized for speed and Norsys offers detailed descriptions of all functions. This partial network is for demonstration purposes. In an actual Bayes net, each facet has multiple indicators connected to it and the actual number of variables included in a Bayes net varies across levels depending on the number of indicators identified. The main problem solving node and its four facets remain in the network throughout all the levels. Any incoming evidence about a student's status on an indicator will update estimates about the facet it belongs to, and the evidence will get propagated through the whole network. This process yields an instantiated BN per student for each level they play.

Now suppose that a player consistently planted iceberg lettuce in front of snapdragons on a given level in PvZ2. The final ratio of iceberg lettuce planted in front of snapdragons to the total number of iceberg lettuces planted is 88%, which belongs to the last, lowest state of the

Figure 8. Bayes net of problem solving (fragment)--prior probabilities



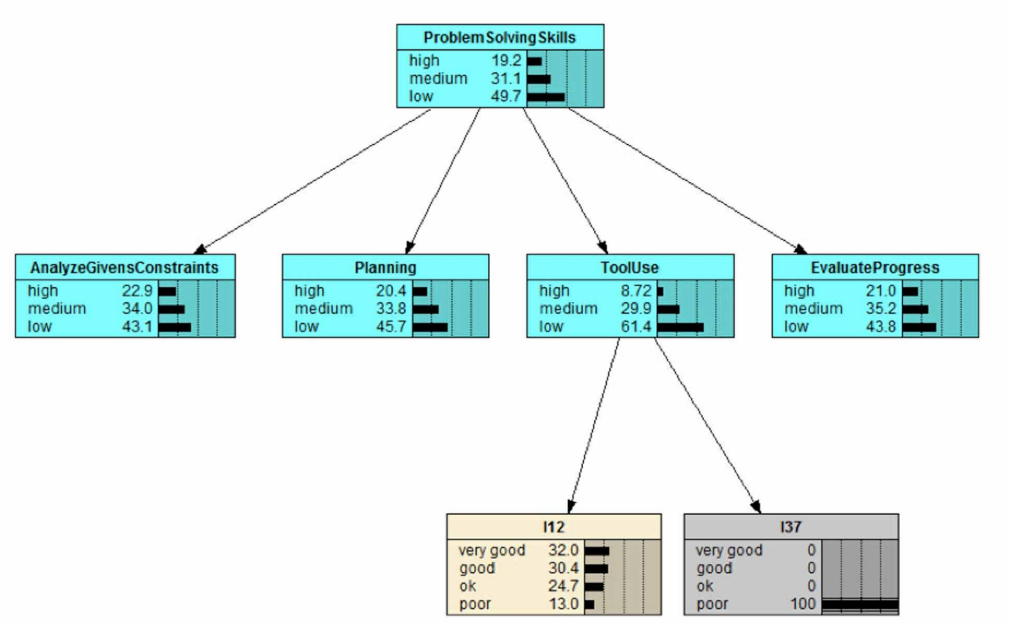
node (i.e., "poor" in indicator #37). This evidence would be entered into the network and then propagated throughout the network (see the updated probability distribution for every node in Figure 9). As a result, the network estimates that the player, at this point in time, is most likely to be low in effective tool use: Pr (use of tools = low | evidence) = 0.61, and thus relatively low in overall problem-solving skill: Pr (problem-solving = low | evidence) = .50.

The player, at some point, will likely become aware of the folly of placing iceberg lettuce in front of a snapdragon. If the player then decides to feed the snapdragon some plant food to boost the snapdragon's power, then this action would suggest that the player understands the function of plant food (indicator #12). Consequently, the power boost effectively wiped out (in a blaze of fire) four zombies on the screen (see the burnt zombies in Figure 10 for the special effect of plant food on snapdragons). This evidence suggests that the player realized that plant food is a scarce resource that should be conserved for critical situations, such as an attack by a large wave of zombies (i.e., at least three zombies). The BN incorporated the evidence and updated the estimates of the player's competencies (see Figure 11). The current probability distribution of the player's level of effective tool use is: Pr (use of tools = low | evidence) = .45, Pr (use of tools = medium | evidence) = .39, Pr (use of tools = high | evidence) = .16. The estimates for the

Table 2. Conditional probability table for indicator #37 in level 7 of the Pirate Seas

Effective/efficient tool use	Very good	Good	Ok	Poor
High	0.53	0.32	0.11	0.04
Medium	0.36	0.36	0.21	0.07
Low	0.19	0.32	0.31	0.18

Figure 9. Evidence of poor use of iceberg lettuce received by the Bayes net



player's problem-solving skill is: Pr (problem-solving skills= high | evidence) = .25, Pr (problem-solving skills = medium | evidence) = .34, and Pr (problem-solving skills = low | evidence) = .41.

When setting up the initial (prior) probabilities in the BN, we assumed that students would have an equal likelihood of being high, medium, or low on problem solving. As more evidence enters the network, the estimates become more accurate and tend to reflect each student's true status on the competency. Evidence is collected dynamically by the game logs. After developing the BNs (one for each level in the game) and integrating them into the game code, we are able to acquire real-time estimates of players' competency levels across the main node (problem-solving skill) and its constituent facets.

To establish construct validity, we tested the correlations among our stealth assessment estimates of problem solving and an external measure—MicroDYN (Wustenberg, Greiff, & Funke, 2012) in a pilot study. We had ten undergraduate students play PvZ 2 for 90 minutes and then complete MicroDYN (30 minutes). MicroDYN is another example of a performance-based assessment. The assessment presents a real-world system in each item, requiring participants to figure out causal relationships among different variables and then manipulate the variables to control the system in specific ways. Towards the goal of testing construct validity, we reduced the probability estimates of the overall problem solving node (e.g., high, medium, and low levels) to a single number. To do this we assigned numeric values +1, 0 and -1 to the three states, and computed the expected value. This Expected A Posteriori (EAP) value can also be expressed as, $P(\theta ij = \text{High}) - P(\theta ij = \text{Low})$, where θij 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 from -1 to 1. The results show that our game-based assessment of problem solving skills is significantly correlated with MicroDYN (r = .74, p = .03) and thus our problem solving stealth assessment is valid. The results need to be further verified with a larger sample size. We are currently running a larger validation study with approximately 50 middle-school students playing PvZ2 for three hours.

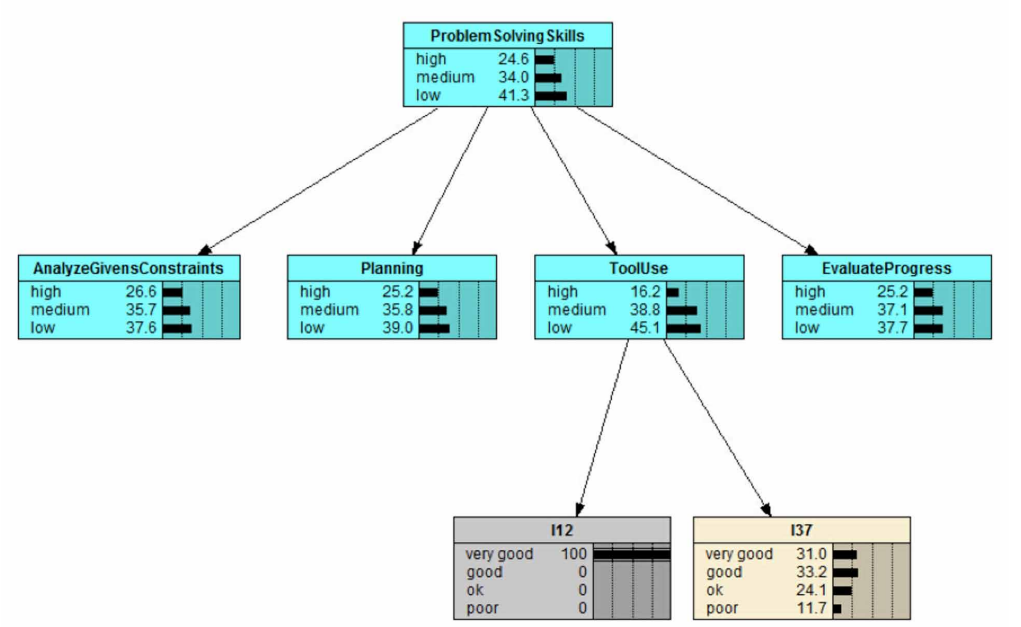
Figure 10. Screen capture of a player using the plant food power boost on Snapdragons

3. DISCUSSION AND CONCLUSION

Stealth assessment represents an innovative and powerful way to ensure the validity of competency measures within a rich and interactive learning or gaming environment. The assessment is woven into the environment such that it becomes invisible to students, which is conducive to eliciting true knowledge and skills. ECD provides a framework for designing stealth assessments that captures far more information related to student competencies than simpler judgments of right or wrong, a single summative score on a test, or responses to self-report queries. When designing stealth assessment, the assessment designer starts by defining the competency model. This ensures that the assessment is firmly grounded in the competency of interest. The designer then determines what indicators from the learning or gaming environment would elicit the evidence. If the designer (or more likely, the design team) needs to create a game or a learning system from scratch, this step should be preceded by defining task models that define the features and constraints of different tasks. Next, the designer works on the evidence model, which involves connecting the indicators and competency model variables statistically.

In addition to being valid and reliable, such ubiquitous and unobtrusive assessments can assist in instructional decision making, such as advancing or remediating students, as warranted. We recommend using BNs as the statistical inference tool in stealth assessments as they enable real-time updates of estimates relative to target competencies, which allow automated assessment machinery and/or assessors to continuously obtain accurate information about the learner. The most up-to-date inferences about students' learning in the environment can be used to identify

Figure 11. The second indicator update involving the good use of plant food



when formative feedback should be provided. Furthermore, as additional evidence is accumulated, the quality of the assessment (in terms of validity and reliability) will invariably improve. Another advantage of BNs over other approaches frequently used in data mining (e.g., item response theory, regression) is that they support multi-dimensionality of the competencies we study. In other words, we adopt BNs because they allow us to model competencies at a fine grain size (i.e., the main competency along with associated facets and perhaps sub-facets) although using BNs requires extra effort in the construction of the competency model (Desmarais & Pu, 2005). BNs are only recently beginning to be used in the area of educational data mining (Baker & Yacef, 2009), particularly given the ability to model latent competencies.

As discussed in this paper, we encountered several challenges in a couple of research projects while developing stealth assessment in games. These challenges taught us some valuable lessons that may help prevent problems for others engaged in similar research. To summarize, the major lessons we learned to date include: (a) the need to select appropriate external measures to validate stealth assessments and examine any learning transfer from the game, and (b) the importance of customizing log files (i.e., capturing just what is needed as evidence to inform the competency model, but not more) to facilitate data analysis and estimation of competency states.

For validation, the scope and format of any external measure must align with the stealth assessment. Otherwise, it would be unclear if the stealth assessment is valid or not. Additionally, it would be difficult to detect any transfer of learning with the selected external measures. Accurate assessment can lead to useful information that will enable us to support student learning across a range of content and areas. Also, quality should be the top criteria in the selection of external measures (i.e., select external measures that are reliable and valid). We encourage the use of performance-based assessments whenever possible, as they have several advantages over self-report surveys or traditional multiple-choice item types. First, performance-based assess-

ment provides an authentic environment where students are expected to apply their knowledge, skills, and other attributes as they engage in a task or construct a response. At the same time, if designed well, performance-based assessment can be less explicit about the true competency being measured (Shute & Ventura, 2013), and thus would suffer less from social desirability effects than typical self-report measures.

Regarding log files, we suggest carefully checking the usability of log files before making the decision to adopt a commercial game because the analysis of log files of student-computer interaction plays a vital role in stealth assessment. Researchers should schedule adequate time to parse the code to determine if they can extract the information they need from the log files. This issue is easier to tackle in homemade games because the game designer can always adjust her code to make it easy to read and include all necessary information.

The major limitation of implementing stealth assessment using ECD is the cost in terms of time and effort, whether it is a commercial or a homemade game. As Almond, Kim, Velasquez, and Shute (2014) discussed, the process usually spans one or more years involving learning scientists, psychometricians, game designers, programmers, and possibly others (e.g., content experts). However, the stealth assessment may be recycled in other games if designed appropriately, using the same theoretical nodes of the target competency and the same statistical models, only coming up with different indicators that represent the specific actions within a particular game.

We hope that these lessons, accompanied by our recommendations for best practices, are useful to other researchers who are interested in developing and using stealth assessment in their research. Stealth assessment is still a relatively new assessment approach. Future research should examine other important knowledge, skills, and personal attributes that can be measured in this manner. We encourage researchers to share their own lessons and best practices for public discussion. Additional research may also examine the extent to which stealth assessment can be scaled to fit other learning environments, allowing for the recycling of previously built models (i.e., competency, evidence, and task models) to make the process more cost-effective. In fact, we have used the same persistence models in various games, such as Physics Playground and Portal 2. We believe that video games provide a meaningful context where players are required to apply various knowledge and skills to succeed. We foresee that as more people employ stealth assessment, more collaboration will happen between teachers, content-experts, game designers, assessment experts, and other important stakeholders to create engaging games that will serve as assessment and learning tools to a large population across the country.

ACKNOWLEDGMENT

We would like to thank Russell Almond for his guidance on the development of BNs, Weinan Zhao for his technical support in parsing log files and running the BNs, and Fengfeng Ke and Matthew Ventura for their work on the Portal 2 project. We would also like to thank the members of the GlassLab team who are supporting our work assessing problem solving in Plants vs. Zombies 2—specifically Jessica Lindl, Liz Kline, Michelle Riconscente, Ben Dapkiewicz, and Michael John. This work was supported by funding from the Bill & Melinda Gates Foundation (Physics Playground research) and the John D. and Catherine T. MacArthur Foundation (Portal 2 research).

REFERENCES

Almond, R. G. (2010). Using Evidence Centered Design to think about assessments. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st Century: Supporting educational needs* (pp. 75–100). New York: Springer-Verlag. doi:10.1007/978-1-4419-6530-1_6

Almond, R. G., DiBello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., & Yan, D. (2001). Models for conditional probability tables in educational assessment. In Jaakkola, T., & Richardson, T. (Eds.), Artificial Intelligence and Statistics 2001 (pp. 137-143). Morgan Kaufmann.

Almond, R. G., Kim, Y. J., Shute, V. J., & Ventura, M. (2013). Debugging the evidence chain. In A. Nicholson & P. Smyth (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Ninth Conference*. Corvallis, OR: AUAI Press.

Almond, R. G., Kim, Y. J., Velesquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement: An Interdisciplinary Perspective*, *12*, 1–33.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Williamson, D. M., & Yan, D. (in press). *Bayesian Networks in Educational Assessment*. New York: Springer.

American Educational Research Association (AERA). American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica*, *129*(3), 387–398. doi:10.1016/j.actpsy.2008.09.005 PMID:18929349

Chermahini, S. A., Hickendorff, M., & Hommel, B. (2012). Development and validity of a Dutch version of the Remote Associates Task: An item-response theory approach. *Thinking Skills and Creativity*, *7*(3), 177–186. doi:10.1016/j.tsc.2012.02.003

Chu, Y., & MacGregor, J. N. (2011). Human performance on insight problem solving: A review. *The Journal of Problem Solving*, *3*(2), 119–150. doi:10.7771/1932-6246.1094

Desmarais, M. C., & Pu, X. (2005). A Bayesian student model without hidden nodes and its comparison with Item Response Theory. *International Journal of Artificial Intelligence in Education*, *15*, 291–323.

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

Hung, W., & Van Eck, R. (2010). Aligning problem solving and gameplay: A model for future research and design. In R. Van Eck (Ed.), Interdisciplinary models and tools for serious games: Emerging concepts and future directions (pp. 227-263). New York: Hershey. doi:10.4018/978-1-61520-719-0.ch010

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kim, Y. J., & Shute, V. J. (in press). Opportunities and challenges in assessing and supporting creativity in video games. To appear. In J. Kaufmann & G. Green (Eds.), *Research frontiers in creativity*. San Diego, CA: Academic Press.

Lanyon, R. I., & Goodstein, L. D. (1997). *Personality assessment* (3rd ed.). New York: Wiley.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A. R., Evans, C., & Vitak, J. (2008). *Teens' gaming experiences are diverse and include significant social interaction and civic engagement.* Retrieved from http://www.pewinternet.org/Reports/2008/Teens-Video- Games-and-Civics.aspx

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. doi:10.1037/h0048850 PMID:14472013

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. doi:10.1207/S15366359MEA0101_02

Partnership for 21st Century Learning. (2012). http://www.p21.org

Paulhaus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17–59). San Diego, CA: Academic Press. doi:10.1016/B978-0-12-590241-0.50006-X

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. doi:10.1037/a0014996 PMID:19254083

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *The British Journal of Medical Psychology*, *19*(1), 137–150. doi:10.1111/j.2044-8341.1941.tb00316.x

Roberts, B. W., Kuncel, N., Shiner, R. N., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socio-economic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345. doi:10.1111/j.1745-6916.2007.00047.x PMID:26151971

Schmitt, N. (1994). Method bias: The importance of theory and measurement. *Journal of Organizational Behavior*, *15*(5), 393–398. doi:10.1002/job.4030150504

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Shute, V. J., & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning? In D. Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 359–387). New York, NY: Routledge Books.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (in press). Advances in the science of assessment. *Educational Assessment*.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning*, *8*(2), 137–161.

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

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, *80*, 58–67. doi:10.1016/j.compedu.2014.08.013

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research*, *106*(6), 423–430. doi:10.1080/00220671.2013.832970

Shute, V. J. & Wang, L. (in press). Measuring problem solving skills in Portal 2. To appear in: *E-learning systems, environments and approaches: Theory and implementation*.

Ventura, M., & Shute, V. J. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572. doi:10.1016/j.chb.2013.06.033

Ventura, M., Shute, V. J., & Small, M. (2014). Assessing persistence in educational games. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), *Design recommendations for intelligent tutoring systems: Instuctional Management* (Vol. 2, pp. 93–101). Orlando, FL: U.S. Army Research Laboratory.

Ventura, M., Shute, V. J., Wright, T., & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology*, *4*, 1–7. doi:10.3389/fpsyg.2013.00852 PMID:24379790

Ventura, M., Shute, V. J., & Zhao, W. (2012). The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, *60*(1), 52–58. doi:10.1016/j.compedu.2012.07.003

Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. *Journal of Experimental Psychology. General*, *110*(2), 169–192. doi:10.1037/0096-3445.110.2.169