Computers in Human Behavior 63 (2016) 106–117

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Measuring problem solving skills via stealth assessment in an engaging video game

Valerie J. Shute [a, *], Lubin Wang [a], Samuel Greiff [b], Weinan Zhao [a], Gregory Moore [a]

[a] Florida State University, Tallahassee, FL, USA
[b] University of Luxembourg, Walferdange, Luxembourg

ARTICLE INFO

Article history:
Received 27 August 2015
Received in revised form
16 April 2016
Accepted 13 May 2016

Keywords:
Problem solving skills
Stealth assessment
Validity

ABSTRACT

We used stealth assessment, embedded in a game called *Use Your Brainz* (a slightly modified version of *Plants vs. Zombies 2*), to measure middle-school students' problem solving skills. We began by developing a problem solving competency model based on a review of the relevant literature. We then identified in-game indicators that would provide evidence about students' levels on various problem-solving facets. Our problem solving model was then implemented in the game via Bayesian networks. To validate the stealth assessment, we collected data from students who played the game-based assessment for three hours and completed two external problem solving measures (i.e., Raven's Progressive Matrices and MicroDYN). Results indicated that the problem solving estimates derived from the game significantly correlated with the external measures, which suggests that our stealth assessment is valid. Our next steps include running a larger validation study and developing tools to help educators interpret the results of the assessment, which will subsequently support the development of problem solving skills.

1. Introduction

Problem-solving skill is generally defined as a person's ability to engage in cognitive processing to understand and resolve problem situations where a method to solve the problem is not immediately available. According to the Organisation for Economic Cooperation and Development (OECD), problem-solving skill also includes the motivation to engage with such situations in order to "achieve one's potential as a constructive and reflective citizen" (OECD, 2014). This important competency is one that we believe should be fully embraced in our education systems. However, according to the recent OECD Report, students in the U.S. rank 15th out of 44 participating countries on the Programme for International Student Assessment (PISA) Problem Solving test.

The Director of Education and Skills at the OECD recently noted that today's 15-year-old students with poor problem-solving skills will develop into tomorrow's adults attempting to find or keep a good job. He recommended a shift towards supporting problem solving skills in school curricula (OECD, 2014). However, one issue with teaching problem solving skills in a classroom context is that the problems presented in formal education tend to be qualitatively different from those encountered in the real world. That is, problems presented in assessment situations in schools are typically clearly defined and structured, whereas problems in real life are often ill-structured. Well-designed digital games offer a viable alternative to assessing and developing complex problem solving skills that are needed to succeed in the real world (Greiff & Funke, 2009; Greiff et al., 2014; Shute & Wang, in press; Shute, Ventura, & Ke, 2015).

U.S. students' mediocre development of problem solving skills is also of concern to American business leaders, who are dissatisfied with college graduates' lack of problem solving skills. A recent survey of business leaders conducted by the Association of American Colleges and Universities indicates that problem solving skills are increasingly desired by American employers, but only 24% of employers report that recently hired American college graduates are able to analyze and solve complex problems at work (Hart Research Associates, 2015). Therefore, developing good problem solving skills is very important to successfully navigating through school, career, and life in general (Bransford & Stein, 1984; Jonassen, 1997).

In this paper, we describe the design, development, and validation of an assessment embedded in a video game to measure the problem solving skills of middle school students. After providing a

* Corresponding author. Florida State University, 1114 West Call Street, Tallahassee, FL, 32306-4453, USA.
E-mail address: vshute@fsu.edu (V.J. Shute).

http://dx.doi.org/10.1016/j.chb.2016.05.047
0747-5632/

brief background on stealth assessment and problem solving skills, we describe the game (*Use Your Brainz*) used to implement our stealth assessment, and discuss why it is a good vehicle for assessing problem solving skills. Afterwards, we present our competency model and in-game indicators (i.e., gameplay evidence) of problem solving, describing how we decided on these indicators and how the indicators are used to collect data about the in-game actions of players. While discussing the indicators, we show how the evidence is inserted into a Bayesian network to produce overall and facet-level estimates of students' problem solving skills (using an example reported in Wang, Shute, & Moore, 2015). We then discuss the results of a validation study, which suggest that our stealth assessment estimates of problem solving skill correlate significantly with external measures of problem solving (i.e., Raven's Progressive Matrices and MicroDYN). We conclude with the next steps in developing the assessment and practical applications of this work.

2. Background

2.1. Stealth assessment

Good games are engaging, and engagement is important for learning (e.g., Arum & Roksa, 2011; Dede, 2009; Taylor & Parsons, 2011). One of the challenges of harnessing the engagement that games can produce for learning is validly and reliably measuring learning in games without disrupting engagement, and then leveraging that information to bolster learning. Over the past eight years, we have examining various ways to embed valid assessments directly into games with a technology called *stealth assessment* (e.g., Shute & Ke, 2012; Shute, 2011; Shute, Leighton, Jang, & Chu, 2016; Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Stealth assessment is grounded in an assessment design framework called evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003). The main purpose of any assessment is to collect information that will allow the assessor to make valid inferences about what people know, what they can do, and to what degree they know or are able to do something (collectively referred to as "competencies" in this paper). ECD is a framework that consists of conceptual and computational models that work together harmoniously. The framework requires one to: (a) define the claims concerning learners' competencies, (b) establish what represents valid evidence of a claim, and (c) determine the kind of tasks or situations that will elicit that evidence.

Stealth assessment complements ECD by determining specific gameplay behaviors that can act as evidence of a claim (specified in the evidence model and referred to as indicators) and linking them to the competency model (Shute & Ventura, 2013). As students interact with each problem (or level) in a game during the solution process, they provide an ongoing stream of performance data, captured in a log file. The performance data is automatically analyzed and scored by the evidence model, then inserted into to the competency model, which statistically updates the claims about relevant competencies in the student model (i.e., the instantiated competency model for each individual). The ECD approach, combined with stealth assessment, provides a framework for developing assessment tasks that are clearly linked to claims about personal competencies via an evidentiary chain (i.e., valid arguments that connect task performance to competency estimates), and thus are valid for their intended purposes. The estimates of competency levels can be used diagnostically and formatively to provide adaptively selected game levels, targeted feedback, and other forms of learning support to students as they continue to engage in gameplay. Given the dynamic nature of stealth assessment, it promises advantages such as measuring learner competencies continually, adjusting task difficulty or challenge in light of learner performance, and providing ongoing feedback.

Some examples of stealth assessment prototypes have been described elsewhere (e.g., Shute et al., 2016) — from systems thinking to creative problem solving to causal reasoning relative to the following games: *Taiga Park* (Shute, Masduki, & Donmez, 2010), *Oblivion* (Shute et al., 2009), and *World of Goo* (Shute & Kim, 2011), respectively. For the game *Physics Playground* (see Shute & Ventura, 2013), three stealth assessments—measuring persistence, creativity, and conceptual physics understanding—were created and evaluated for validity and reliability, student learning, and enjoyment (see Shute, Ventura, & Kim, 2013). The stealth assessments correlated with associated externally validated measures for construct validity, and demonstrated reliabilities around 0.85 (i.e., using intraclass correlations among the in-game measures such as the number of gold trophies received). Moreover, 167 middle school students significantly improved on an external physics test (administered before and after gameplay) despite no instructional support relative to the physics content in the game. Students also enjoyed playing the game (reporting a mean of 4 on a 5-point scale, where 1 = strongly dislike and 5 = strongly like).

In summary, some of the benefits of employing stealth assessment include: providing assessments in engaging and authentic environments, reducing or eliminating test anxiety (which can hamper validity), measuring competencies continuously over time, providing ongoing feedback to support learning, and adjusting the difficulty of the learning/gaming environment in response to a person's current level of understanding or skill at various grain sizes (i.e., overall and at the sub-skill level).

Next, we review our focal competency—problem solving skill—in terms of its underlying conceptualization, and discuss the natural fit between this construct and particular video games (i.e., action, puzzle solving, and strategy games).

2.2. Problem solving skills

Problem solving has been studied by researchers for many decades (e.g., Anderson, 1980; Gagné, 1959; Jonassen, 2003; Mayer & Wittrock, 2006; Newell & Shaw, 1958) and is seen as one of the most important cognitive skills in any profession, as well as in everyday life (Jonassen, 2003). Mayer and Wittrock (1996, 2006) identified several characteristics of problem solving: (a) it is a cognitive process; (b) it is goal directed; and (c) the complexity (and hence difficulty) of the problem depends on one's current knowledge and skills.

In 1984, Bransford and Stein integrated the collection of problem-solving research at that time and came up with the IDEAL problem solving model. Each letter of IDEAL stands for an important part of the problem solving process: *Identify* problems and opportunities; *define* alternative goals; *explore* possible strategies; *anticipate* outcomes and act on the strategies; and *look* back and *learn.* Gick (1986) presented a simplified model of the problem-solving process, which included constructing a representation, searching for a solution, implementing the solution, and monitoring the solution (also see the PISA conceptualization of individual and interactive problem solving, OECD, 2014). More recent research suggests that there are two overarching facets of problem-solving skills that can be empirically distinguished and that usually collate several of the more narrow processes mentioned above: rule (or knowledge) acquisition, and rule (or knowledge) application (Schweizer, Wüstenberg, & Greiff, 2013; Wüstenberg, Greiff, & Funke, 2012). "Rules" are the principles that govern the procedures, conduct, or actions in a problem-solving context. Rule acquisition (or identification) involves acquiring knowledge of the

problem-solving environment, whereas rule application involves controlling the environment by applying that knowledge. This is how we conceptualize problem solving skills in this project.

Having defined our working definition of problem solving skill, the next logical question is whether or not these skills can be improved with practice. Polya (1945) noted that people are not born with problem-solving skills. Rather, people cultivate these skills when they have opportunities to solve problems. Furthermore, many researchers have suggested that a vital purpose of education should be to teach people to become better problem solvers (Anderson, 1980; Ruscio & Amabile, 1999). However, there is a gap between problems in formal education and those that exist in real life. Jonassen (2000) noted that the problems students encounter in school are mostly well-defined, which contrast with real-world problems that tend to be messy, with multiple possible solutions. Moreover, many problem-solving strategies that are taught in school entail a "cookbook" type of memorization and result in functional fixedness, which can impede students' ability to solve novel problems and develop their own knowledge-seeking skills (Jonassen, Marra, & Palmer, 2004). This is where well-designed digital games become relevant. Such games consist of a set of goals and complicated scenarios that require the player to generate new knowledge and skills in order to advance through the game. Researchers (e.g., Shute et al., 2009; Van Eck, 2006) have argued that playing well-designed games (such as action, puzzle solving, and strategy genres) can promote problem-solving skills because games require a continual interaction between the player and the game, often in the context of solving stimulating problems that increase in difficulty.

Empirical research examining the effects of video games on problem-solving skills is currently scarce. In one study though, Buelow, Okdie, and Cooper (2015) reported that students in their video game condition demonstrated significantly better problem solving skills (as assessed by the Wisconsin Card Sorting Task) compared to a no-game control group. The game used in the study by Hou and Li (2014) was designed following the problem-based gaming model (Kiili, 2007) that encouraged students to utilize different problem solving strategies. However, problem solving skills were not directly assessed in their study. Our present research is intended to help fill this gap in the literature.

3. Methods

3.1. Participants

Our sample consisted of 55 7th grade students enrolled at a middle school in suburban Illinois. They each played the game *Use Your Brainz* on an iPad for about three hours across three consecutive days. On the fourth day, all students completed two external tests of problem solving skill: (a) Raven's Progressive Matrices (Raven, 1941, 2000), which measures reasoning and simple problem solving skills, and (b) MicroDYN (Wüstenberg et al., 2012), which measures complex problem solving skills. Each student also completed a demographic questionnaire concerning his or her age, gender, gaming history, and so on. Together, these assessments took about one hour to complete.

Seventh grade students were selected because of the alignment between the problem solving content and the math content in the Common Core State Standards at that grade level (e.g., MP1: *Make sense of problems and persevere in solving them*, and MP5: *Use appropriate tools strategically*). Among the 55 participants, one student's gameplay data was missing, five students did not take the Raven's Progressive Matrices test, and two students did not complete the MicroDYN test. After we excluded the missing data, we had complete data from 47 students (20 male, 27 female).

3.2. The game - Use Your Brainz

The game we employed was a slightly modified version of the popular game *Plants vs. Zombies 2* (Popcap Games and Electronic Arts) called *Use Your Brainz* (UYB). This game served as the vehicle for our problem solving stealth assessment. In the game, players must plant a variety of special plants on their lawn to prevent zombies from reaching their house. Each of the plants has different attributes. For example, some plants (offensive ones) attack zombies directly, while other plants (defensive ones) slow down zombies to give the player more time to attack the zombies. A few plants generate "sun," an in-game resource needed to produce more plants. The challenge of the game comes from determining which plants to use and where to place them in order to defeat all the zombies in each level of the game.

We chose *UYB* as our assessment environment for two reasons. First, we were able to modify the game because of our working relationship with the Glasslab. Glasslab was able to obtain access to the source code for *Plants vs. Zombies 2* and make direct changes to the game as needed (e.g., specify and collect particular information in the log files). This is important because it allows us to build the stealth assessment directly into the game itself. Second, *UYB* requires players to apply problem solving skills as an integral part of gameplay. Thus, our stealth assessment is able to collect ongoing data relevant to problem solving skills directly from students' gameplay.

3.3. Problem solving model

We developed a problem solving competency model based on a comprehensive review of the problem solving literature. We divided problem solving skill into four primary facets: (a) analyzing givens and constraints, (b) planning a solution pathway, (c) using tools and resources effectively and efficiently, and (d) monitoring and evaluating progress. The first facet maps to "rule acquisition" and the remaining facets map to "rule application." After defining the facets of problem solving, we identified relevant in-game indicators for each of the four facets (see next section for details).

The rubrics for scoring each indicator during gameplay and the statistical links between the indicators and the competency model variables comprise the evidence model. The competency and evidence models are implemented together in Bayesian networks. We created a unique Bayes net for each game level (42 total levels comprising two different worlds within the game UYB: *Ancient Egypt* and *Pirate Seas*) because most of the indicators do not apply in every level and simple networks make computations more efficient.

In the Bayes nets, the overall problem solving variable, each of the four facets, and the associated indicators are nodes that influence each other. Each node has multiple potential states and a probability distribution that defines the likely true state of the variable. Bayes nets accumulate data from the indicators and propagate this data throughout the network by updating the probability distributions. In this way, the indicators dynamically influence the estimates of a student's problem solving skill, overall and on each of the four problem solving facets.

3.4. Indicators of problem solving

In line with the stealth assessment process, we defined indicators for each of the four facets of problem solving by identifying observable, in-game actions that would provide evidence per facet. This was an iterative process that began by brainstorming a large list of potential indicators derived from playing through the game multiple times and watching a range of solutions to some difficult

levels posted on YouTube.

After listing all potential indicators, we evaluated each one for (a) *relevance* to its associated facet(s), and (b) *feasibility* of being implemented in the game. We removed indicators that were not closely related to the facets or were too difficult or vague to implement. We repeated this process of adding, evaluating, and deleting indicators until we were satisfied with the list. We ended up with 32 indicators for our game-based assessment: 7 for analyzing givens and constraints, 7 for planning a solution pathway, 14 for using tools and resources effectively, and 4 for monitoring and evaluating progress. Examples of indicators for each facet are shown in Table 1.

3.5. Estimating problem solving skill with Bayes nets

Once the set of observable variables was determined (which included both positive and negative indicators), our next step was to figure out how to score the observables and establish acceptable statistical relationships between each observable and the associated levels of the competency model variables. Scoring rules were based on the collection of relevant instances of observables and then a classification into discrete categories, such as yes/no (a student did, or did not do some action in the level), or poor/ok/good/very good (depending on the quality of the actions). We constructed Bayesian networks (BNs) to accumulate the incoming data and update beliefs relative to the competency levels.

A BN graphically reveals the conditional dependencies that exist between the various variables in the network. It consists of both competency model variables (i.e., problem solving and its four facets) and the associated observables (indicators) that are statistically linked to the facets. As mentioned, we constructed a separate BN for each level in the game because the observables change across levels, and levels differ in terms of difficulty. Estimates related to a player's problem solving skill are updated as ongoing evidence accrues from his or her interactions with the game. For example, a facet of problem solving is the ability of a player to *use tools effectively and efficiently*. One of the dozens of plants in the game is called Iceberg Lettuce, which is used to freeze an incoming zombie temporarily, thus delaying the zombie's attack (see the top row, right side of Fig. 1 for the result of a zombie coming in contact with Iceberg Lettuce). Another plant in the game is the Snapdragon, which exhales fire. The Snapdragon is planted in order to burn approaching zombies. Both of these plants (and many others) serve to hinder the onslaught of zombies, and are thus considered valuable resources or tools, if used properly. But 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 end up melting the ice from the Iceberg Lettuce immediately, rendering it useless. If a player makes this error, 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.

Table 2 displays the communication between the log files and relevant BN nodes. The first row describes indicator #37: *Player plants Iceberg Lettuce within range of a Snapdragon attack (3x3 space in front of a snapdragon).* Any time that a player plants an Iceberg Lettuce in the game, the scripts that run in the game logging system command a check for a Snapdragon in proximal tiles. At the end of a level, a proportion is calculated involving the number of Iceberg Lettuces planted in the range of a snapdragon divided by the total number of Iceberg Lettuces planted. Because this is an undesirable action (inversely coded), a *lower* ratio represents better performance. For this indicator, performance is categorized into one of four levels, ranging from poor to very good. Each level comprises 25% of the distribution. If the ratio falls within [0, 0.25], then this evidence corresponds to the "very good" state in the BN node (i.e., indicator #37). Similarly, if the ratio falls within [0.26, 0.5], it corresponds to the "good" state; [0.51, 0.75] reflects the "ok" state; and [0.76, 1] corresponds to the "poor" state of the node in the network.

The statistical relationships involving indicator #37 and its associated competency variable (effective/efficient tool use) are defined by the likelihood in a conditional probability table (CPT). For instance, Table 3 shows the conditional probability table for indicator #37 in level 7 of the Pirate Seas world in UYB. For example, the value of 0.53 in the first cell indicates that if the player is (theoretically) high on effective/efficient tool use, the likelihood is 0.53 that he or she will fall in the "very good" state of indicator #37. After performance data (evidence) about a student's observed results on indicator #37 arrives from the log file, the estimates on his or her ability to use tools effectively will be updated based on Bayes theorem. We configured the distributions of the 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, 2010; Almond, Mislevy, Steinberg, Yan, & Williamson, 2015).

We set the discrimination estimate for indicator #37 to 0.3 (i.e., low). According to Almond, Kim, Shute, and Ventura (2013), discrimination in game-based assessment is expected to be fairly low because of the many confounds involved, like prior gameplay experience. We set the difficulties for the best (i.e., very good) to worst (i.e., poor) states to 0, −1, −2, and −3, respectively (i.e., this is a fairly easy indicator/task to learn). These parameters were initially determined by a learning scientist, two Plants vs. Zombies game experts, and a psychometrician. The CPTs were later calibrated with empirical data collected from a pilot study using the game. The values for the discrimination and difficulty parameters per indicator and per level were documented within an augmented Q-matrix for possible future adjustment (Almond, 2010). In our Q-matrix, the rows represent the indicators that are relevant in each level of the game, and the columns are the four problem solving facets.

Fig. 2 shows the prior probabilities for a *fragment* of the problem solving Bayes net—the overall problem solving node along with its four facets and two example indicators related to effective tool use

Table 1
Examples of indicators for each problem solving facet.

Facets	Examples of Indicators
Analyzing Givens & Constraints	• Plants >3 Sunflowers before the second wave of zombies arrives • Selects plants off the conveyor belt before it becomes full
Planning a Solution Pathway	• Places sun producers in the back, offensive plants in the middle, and defensive plants up front • Plants Twin Sunflowers or uses plant food on (Twin) Sunflowers in levels that require the production of X sun
Using Tools and Resources Effectively	• Uses plant food when there are > 5 zombies in the yard or zombies are getting close to the house (within 2 squares) • Damages > 3 zombies when firing a Coconut Cannon
Monitoring and Evaluating Progress	• Shovels Sunflowers in the back and replaces them with offensive plants when the ratio of zombies to plants exceeds 2:1

Fig. 1. Iceberg Lettuce in *UYB.*

Table 2
The communication between log files and relevant Bayes net nodes (facets).

Facet	Indicator #	Relevant Indicators	Telemetry event(s) used	Tech implementation specifications
Effective/ efficient tool use	37	Player plants iceberg lettuce within range of a snapdragon attack (3 × 3 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"

Table 3
Conditional probability table for indicator #37 in level 7 of the Pirate Seas world.

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

(i.e., Indicators #12 and #37). We use the program *Netica* (by Norsys Software Corporation) to construct and compile the network 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.

The partial network shown in Fig. 2 is for illustration purposes. In our operational BNs, each facet is connected to multiple indicators—a subset of the 32 total indicators. The number of variables included in a BN varies across levels given the differential nature of the levels and thus applicable indicators. That said, the main problem solving node and its four facets remain in the network throughout all levels. All incoming evidence about a student's status on an indicator serves to update the estimates about its linked facet(s). In turn, the evidence is propagated throughout the entire network. This process yields an updated BN per student for each level they play.

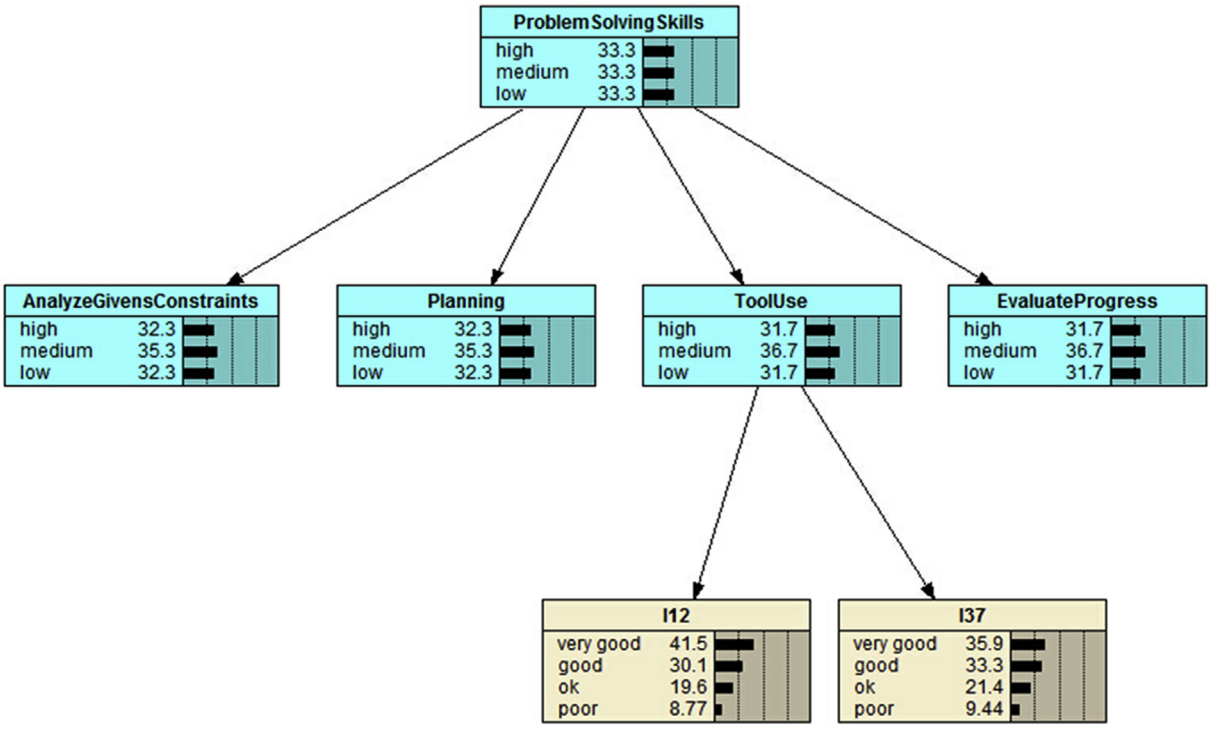

Fig. 2. Fragment of the problem solving Bayes net—prior probabilities. (adapted from Wang, Shute, & Moore, 2015).

Continuing with the example of a player who consistently planted Iceberg Lettuce in front of a Snapdragon within a given level in the game, suppose that the final ratio of Iceberg Lettuce planted in front of Snapdragons to the total number of Iceberg Lettuces planted was 88%. This value would be categorized into the lowest state of the node (i.e., "poor" in indicator #37), and the evidence would be inserted into and propagated throughout the BN (see the updated probability distribution for each node in Fig. 3). The network – at this point in time – would estimate that the player 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) = 0.50.

During a subsequent trial (e.g., either a new level or a replay of a failed level), suppose that the player correctly identified and rectified the blunder of planting Iceberg Lettuce close to Snapdragons. Further, he or she is now focused on expanding the power of Snapdragons. Feeding the Snapdragon (or any plant) some plant food serves to substantially boost the plant's power. This becomes especially important when there is a wave of (or at least three) zombies on the screen. The Snapdragon's power boost can effectively wipe out (in a blaze of fire) multiple zombies at once (see the burnt zombies in Fig. 4 for the special effect of plant food on Snapdragons). However plant food is a very limited resource, and if it is used prematurely, it is wasteful. Using plant food only when there are multiple zombies present would suggest that the player understands the function of plant food and realizes that plant food is a scarce resource that should be conserved for critical situations, such as an attack by a large wave of zombies (indicator #12—see the last column in Table 2 for the associated scoring rule).

The BN incorporates this new evidence and updates the estimates of the player's competencies (see Fig. 5) from the last states shown in Fig. 3. The probability distribution of the player's level of effective tool use is: Pr (use of tools = low) = 0.45, which has decreased (i.e., the player improved) from the last state based on the prior performance relative to indicator #37. Pr (use of tools = medium) = 0.39, which has increased from the last state (again, the player is doing better); Pr (use of tools = high) = 0.16, which has also increased from before. The probability distribution for the player's overall problem-solving skill shows the same pattern of improvement as with tool use.

We defined the initial (prior) probabilities in the BNs based on the assumption that students would have an equal likelihood of being high, medium, or low in terms of their problem solving skills. This means that each student starts with the same initial student model. However, as evidence of the individual student's performance enters the network, the estimates become progressively more accurate – reflecting the student's true status on the competency. This evidence is collected in an ongoing manner 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 each player's competency levels across the main node (problem-solving skill) and its constituent facets.

We now describe the external measures used to establish validity of our in-game estimates of problem-solving skill.

3.6. External measures

Raven's Progressive Matrices (RPM; Raven, 1941) is a test that examines subjects' ability to make inferences based on given information. Typically, subjects are presented with a matrix of eight patterns and one missing slot, arranged in three rows and three columns (see Fig. 6). We selected 12 *RPM* items of increasing difficulty levels for this research.

MicroDYN (Wüstenberg et al., 2012) is a simulation system that tests subjects' ability to acquire and apply information in complex problem solving environments (see Fig. 7 for an example item).

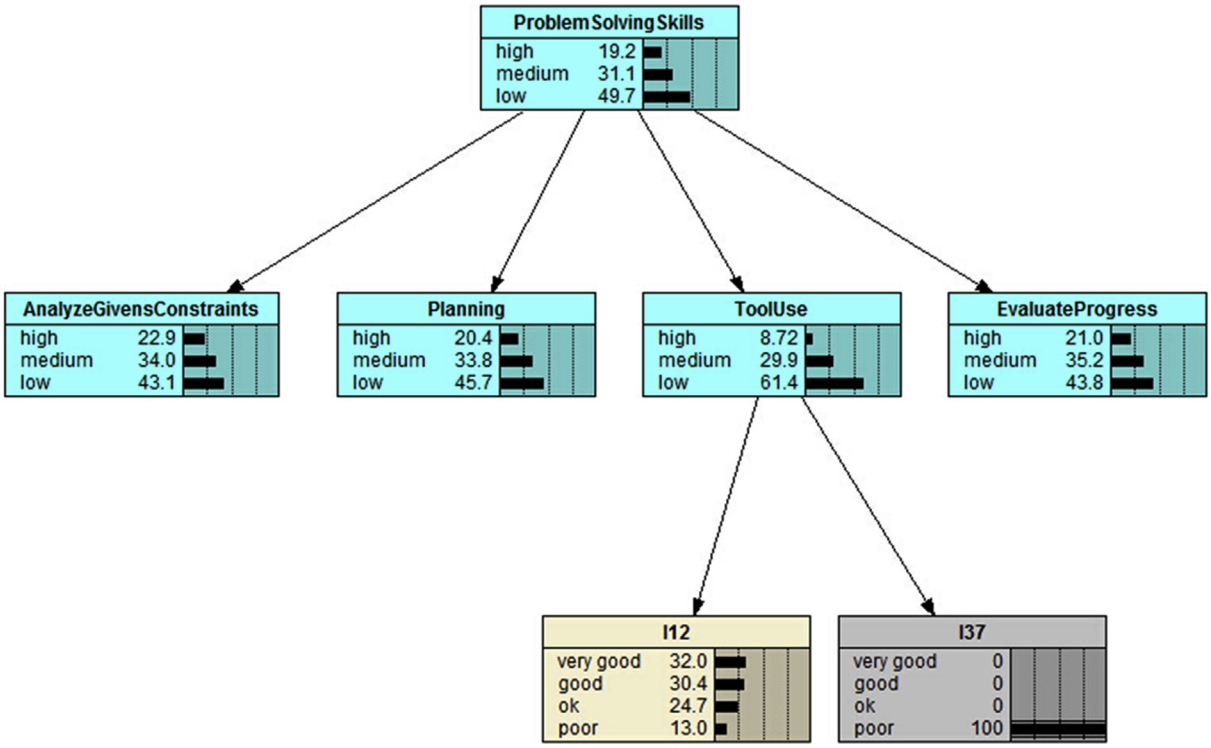

Fig. 3. Evidence of poor use of Iceberg Lettuce received by the Bayes net. (adapted from Wang, Shute, & Moore, 2015).

Fig. 4. Screen capture of a player using the plant food power boost on Snapdragons.

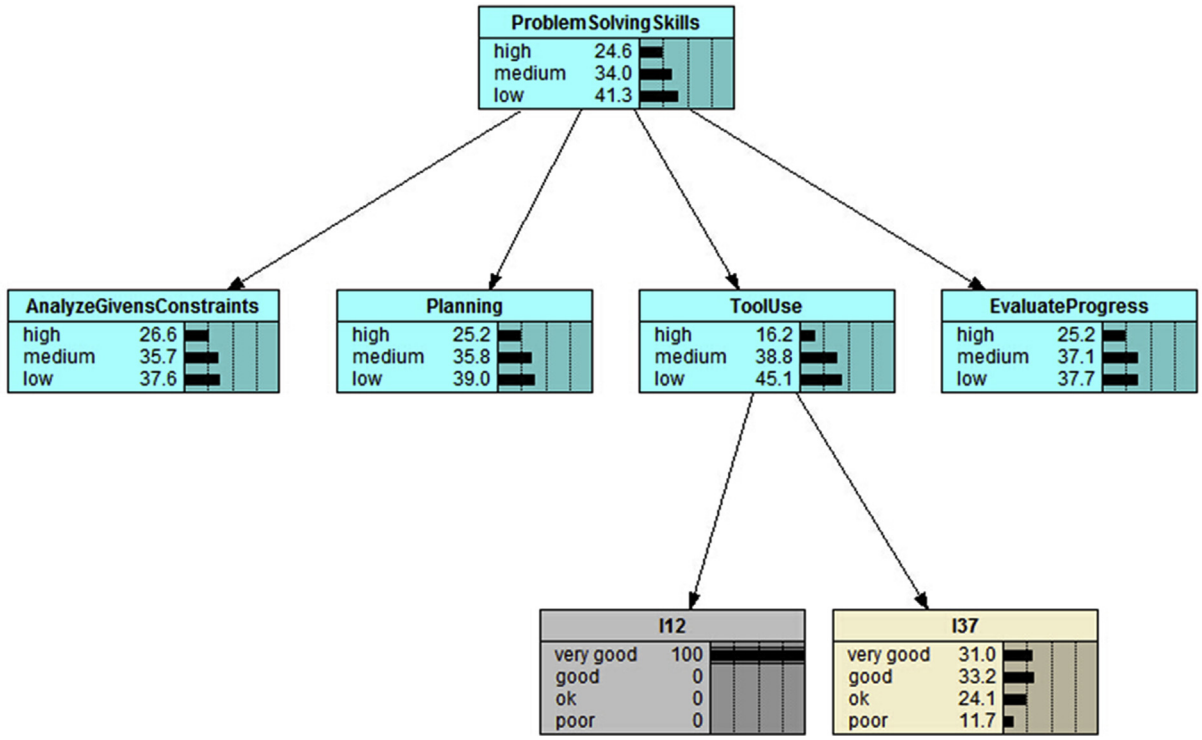

Fig. 5. BN update when plant food is effectively used. (adapted from Wang, Shute, & Moore, 2015).

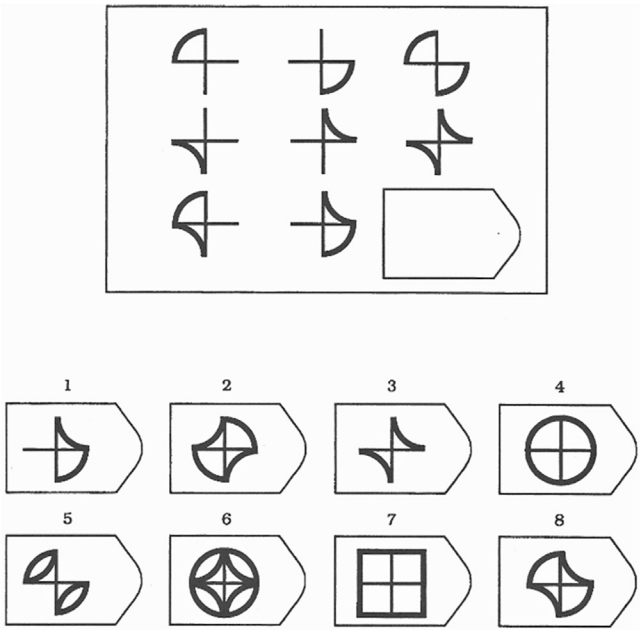

Fig. 6. An example Raven's Progressive Matrices item.

Unlike Raven's Progressive Matrices, not all information is given at the outset in *MicroDYN*. Subjects need to interact dynamically with the system, make decisions, and take actions based on feedback from the environment. There are two phases of the test: (a) depicting the right relationships between the input and output variables (i.e., rule/knowledge acquisition), and (b) reaching the given target values (i.e., rule/knowledge application). Results on MicroDYN were thus divided into two parts relative to rule acquisition and rule application. We computed MicroDYN overall as the average of rule acquisition and rule application. The test requires minimal prior knowledge to complete. The variables in the scenarios were either labeled without deep semantic meaning (e.g., button A) or with fictitious names (e.g., sungrass for a flower). For a thorough overview of MicroDYN and its theoretical background, psychometric properties and validity, see Schweizer et al. (2013) and Wüstenberg et al. (2012).

4. Results

The stealth assessment estimates consisted of three probabilities (i.e., high, medium, and low). However, to analyze these estimates, we needed to reduce the estimates of the overall problem solving node and each of the four facets to a single number. To do this, we assigned numeric values + 1, 0 and −1 to the three states (high, medium, and low), 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 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 from −1 to 1. Table 4 displays the descriptive statistics for the major measures in this study. To keep the scale consistent across measures, we converted all variables to a 0–1 scale.

Students' stealth assessment estimates of their overall problem solving skill, planning a solution, tool usage, and their Raven's Progressive Matrices test scores were generally normally distributed. The distribution of the *analyzing givens and constraints* facet was negatively skewed. This, coupled with its relatively high mean score, suggests that it was the easiest of the four problem solving facets for the students. The distribution for the evaluating progress facet, as well as the three MicroDYN measures were all are positively skewed, indicating that these measures were relatively

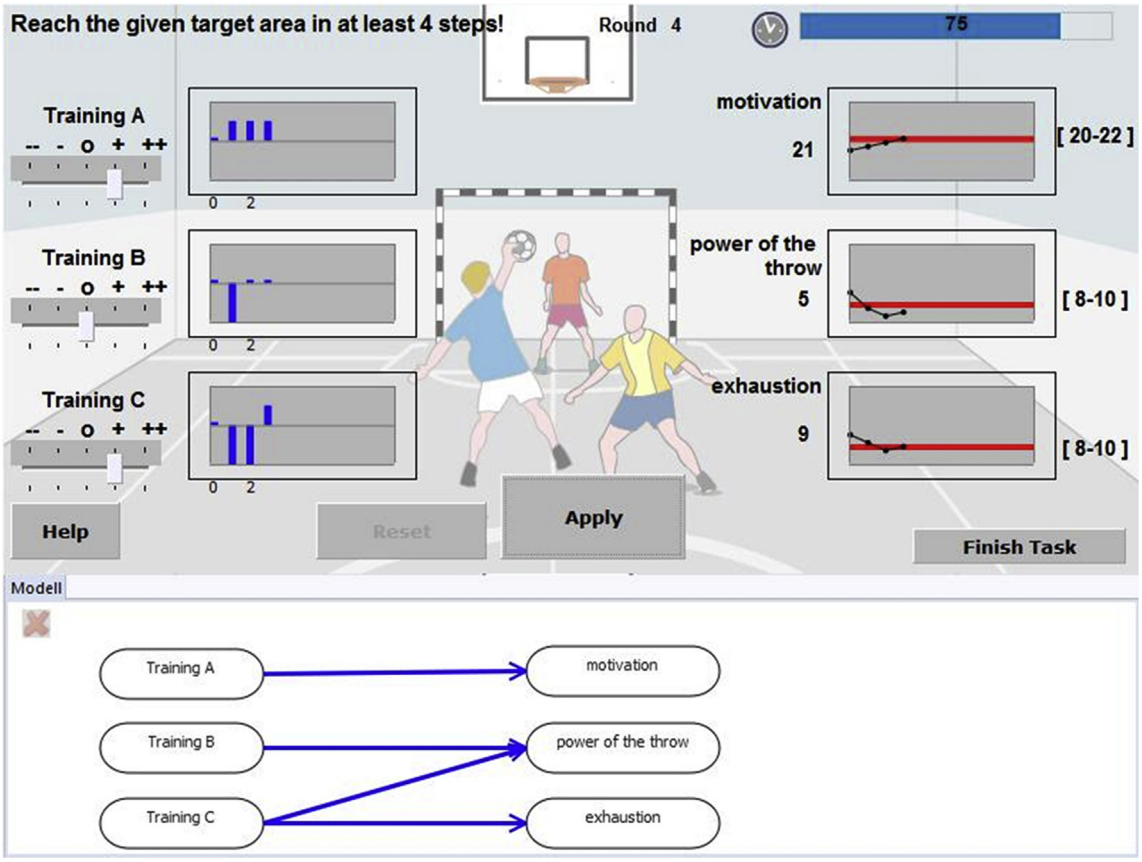

Fig. 7. A sample MicroDYN item.
(from Wüstenberg et al., 2012).

Table 4
Descriptive statistics of the main measures used in the study.

Variable	Mean	SD
Problem solving	0.55	0.31
Analyzing givens & constraints	0.72	0.32
Planning solution	0.63	0.30
Tool usage	0.50	0.27
Evaluating progress	0.16	0.22
Raven's	0.41	0.18
MicroDYN overall	0.14	0.14
MicroDYN rule acquisition	0.10	0.17
MicroDYN rule application	0.19	0.19

difficult for the students.

We conducted internal consistency tests on various measures. For the problem solving stealth assessment, Cronbach's $\alpha = 0.76$, which indicates good internal reliability among the four problem solving facets. However, Cronbach's $\alpha = 0.67$ for rule acquisition, and Cronbach's $\alpha = 0.43$ for rule application relative to MicroDYN, were less than ideal. This may be due to middle school students' difficulty with the tasks (e.g., drawing the model correctly). Future research using this external measure for the same population (i.e., middle school students) should reduce the difficulty of MicroDYN's problem solving tasks. Note that in other reported MicroDYN studies with older samples, reliabilities have been acceptable (e.g., Greiff & Wüstenberg, 2014).

To establish convergent validity, we tested the correlations among our stealth assessment estimates of problem solving skill and the scores from our two external measures. Despite the relatively small sample size ($n = 47$), the results showed that our game-based assessment of problem solving skills significantly correlated with both Raven's ($r = 0.40$, $p < 0.01$) and MicroDYN ($r = 0.41$, $p < 0.01$). Thus our problem solving stealth assessment appears to be valid.

We also found that the estimates for each of the four problem-solving facets were inter-correlated (see Fig. 8 for all manifest correlations between the measures). This is not surprising given that the different indicators per facet come from the same task

	2	3	4	5	6	7	8	9
External Measures:								
1—Ravens	.41**	.43**	.52**	.40**	.42**	.36*	.25	.04
2—Rule Acq, MicroDYN (F1)		.31*	.82**	.24	.24	.25	.08	.04
3—Rule App, MicroDYN (F2)			.80**	.43**	.29*	.37*	.34*	.35**
4—MicroDYN (overall)				.41**	.33*	.38**	.25	.22
In-game Estimates:								
5—Problem solving (overall)					.79**	.98**	.62**	.53**
6—Analyzing givens (F1)						.80**	.24	.31*
7—Planning solution (F2)							.50**	.48**
8—Tool usage (F3)								.23
9—Evaluating progress (F4)								

** p < .01 level; * p < .05; F = facet

Fig. 8. Correlations among the problem solving measures.

(level) in the game, so it is difficult to completely separate relations among indicators. What is a little surprising is the correlation between the overall Problem Solving estimate and the Planning a Solution facet ($r = 0.98$), almost indicating identity. First, note that planning a solution is conceptually the most relevant and broadest of the problem solving facets, thus it makes sense that there would be a strong relation to the overall score. Second, regression results, described in the next section, differ for the four problem solving facets (e.g., their relation to MicroDYN). This indicates that it makes sense to distinguish these facets. Finally, we are also using the data from this validation study to further improve our Bayesian models of problem solving by using machine learning in Netica. As suggested by the model, students' problem solving skill was mainly informed by indicators associated with planning a solution. We need to identify additional strong indicators associated with other facets that can influence the overall problem solving estimate. These results also must be verified with a larger sample size.

To further examine the relative contribution of the two external measures (i.e., Raven's and MicroDYN scores) to the problem solving estimate, we computed a multiple regression analysis with Raven's entered as the first independent variable and MicroDYN second. Results showed that, with both variables in the equation, $R^2 = 0.22$. Most of the problem solving variance was explained by MicroDYN (i.e., Raven's $b = 0.25$; $p = 0.11$; MicroDYN $b = 0.28$, $p = 0.08$). Specifically, MicroDYN explained an additional 5.7% ($p = 0.08$) of the unique variance of our stealth assessment measure beyond Raven's.

Given MicroDYN's overall relation to the stealth assessment measure, we then tested the two components of MicroDYN, computing separate multiple regression analyses to predict the overall problem solving estimate and each of the four facets. When rule acquisition was entered as the first predictor, adding rule application significantly improved the model predicting the overall problem solving estimate ($R^2_{\text{change}} = 0.14$, $p < 0.05$), planning ($R^2_{\text{change}} = 0.09$, $p < 0.05$), tool usage ($R^2_{\text{change}} = 0.11$, $p < 0.05$), and evaluating progress ($R^2_{\text{change}} = 0.11$, $p < 0.05$). However, the same model did not significantly predict analyzing givens ($R^2_{\text{change}} = 0.05$, $p = 0.12$). When rule application was entered as the first predictor, adding rule acquisition did not contribute anything towards explaining the dependent variables (all $p > 0.05$). This confirmed our assumption that the second aspect of MicroDYN, which examines knowledge application, is more relevant to our problem solving stealth assessment, especially planning a solution, using tools effectively, and evaluating progress.

5. Summary and discussion

Today's students are expected to develop 21st century skills such as problem solving (Partnership for 21st Century Learning, 2012). Such 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 their problem solving skills. Such assessments can help educators determine not only students' current levels of this competency, but also students' strengths and weaknesses on particular facets of problem solving. This information can assist educators in supporting their students' development of problem solving skills. However, traditional formats for assessing learning and achievement, such as multiple-choice tests and short-answer questions, often measure superficial skills and are deprived of the context in which knowledge and skills are applied (Shute et al., 2016). An ongoing problem in education thus involves identifying and/or creating more authentic and valid ways to assess students on complex competencies like problem solving. Stealth assessment (Shute, 2011) was proposed as a promising method for assessing such a complex skill. This research builds on this idea by contributing an example of how to create and embed high-quality stealth assessment into a video game.

Based on our analyses, MicroDYN was generally a stronger predictor for the in-game measures of problem solving than Raven's Progressive Matrices. We believe this is because of the conceptual similarities between MicroDYN and UYB, both of which require complex and dynamic problem solving. In addition, our problem solving estimate was more strongly related to MicroDYN's rule-application facet ($r = 0.43$, $p < 0.01$) than rule-acquisition ($r = 0.24$, NS). For Raven's Progressive Matrices, all information is given to the student at the outset, and therefore may not be suitable to test one's ability to accumulate and apply new information continuously. Furthermore, analyses showed that Raven's was not correlated with "tool usage" or "evaluating progress" (see Fig. 8). This brings up another important issue concerning the selection of an appropriate assessment for use in validating a newly created stealth assessment—the alignment between the two. From this perspective, MicroDYN was a more appropriate external measure relative to our particular stealth assessment measure. However, we still need to do more work in order to sufficiently understand the relationships between problem solving, our stealth assessment, and the external measures. It may be worthwhile to examine other competencies and skills that may be related to problem solving. For example, creativity and problem solving are both 21st century skills, and creative students may also tend to be good problem solvers. Examining these types of relationships will in turn allow us to build better models of problem solving.

5.1. Limitations

There are several methodological issues with this validation study. First, the sample of students was small (47 students) and we adjusted the parameters of Bayes nets based on results from the limited sample. As a result, the Bayes nets may suffer from an overfitting problem. This means that the Bayes nets may not fit another

	2	3	4	5	6	7	8	9
External measures:								
1—Ravens	0.41**	0.43**	0.52**	0.40**	0.42**	0.36*	0.25	0.04
2—Rule Acq, MicroDYN (F1)		0.31*	0.82**	0.24	0.24	0.25	0.08	0.04
3—Rule App, MicroDYN (F2)			0.80**	0.43**	0.29*	0.37*	0.34*	0.35**
4—MicroDYN (overall)				0.41**	0.33*	0.38**	0.25	0.22
In-game estimates:								
5—Problem solving (overall)					0.79**	0.98**	0.62**	0.53**
6—Analyzing givens (F1)						0.80**	0.24	0.31*
7—Planning solution (F2)							0.50**	0.48**
8—Tool usage (F3)								0.23
9—Evaluating progress (F4)								

**p < 0.01 level; *p < 0.05; F = facet.

sample as well as they fit the data collected from our small sample. Second, our sample was middle school students. While middle school students do enjoy playing UYB, they are not as good at MicroDYN as older students (high school and university; see Greiff & Wüstenberg, 2014), which was revealed by the relatively low scores on that test. Finally, the participants played the game in a rather contrived (classroom) setting, where they had to start and stop at the same time. To improve the validity and reliability of the stealth assessment, players need to engage in gameplay for longer periods of time spread across multiple sessions.

5.2. Next steps

We are currently examining the results from our validation study to see what additional adjustments need to be made to our Bayes nets (e.g., a level we initially considered to be "difficult" may not be as hard as we thought). Our long term goal is to implement the UYB game-based assessment in middle school classrooms to help educators measure and improve students' problem solving abilities. As part of this effort, we teamed with GlassLab to design a report for their GlassLab Games dashboard that allows educators to easily interpret the results of the assessment—overall and at the individual facet level. There are four different states a student can be in for overall problem solving skill and its four facets: green/mastered (estimated as "high"), yellow/approaching mastery (estimated as "medium"), red/not mastered (estimated as "low), and grey (need more gameplay data). The decision for color per competency node is figured as follows. A node is grey if any pair of probabilities is too close (i.e., <0.15 apart). If the node is not grey, we calculate EAP values. The color of the node is: (a) Green, if EAP falls in [0.34, 1]; (b) Yellow, if EAP falls in [-0.34, 0.33]; and (c) Red, if EAP falls in [−1, −0.33].

This focus on the validity and practicality of our game-based problem solving assessment makes it much more likely that the assessment will be both accurate and useful in classroom settings. Students can be assessed on problem solving, a key cognitive skill, in an engaging environment that presents rich problem solving situations and can parse complex patterns of students actions. Teachers get a valuable tool that will allow them to pinpoint students' abilities in various aspects of problem solving and, in turn, help each student improve their problem solving skills. For example, suppose that a student's overall problem solving skill was estimated as "medium" which was a function of the student being estimated as "high" relative to analyzing givens, but "low" relative to monitoring and evaluating progress. The teacher has something concrete on which to focus support—perhaps suggesting to the student different reflective practices to support her evaluation of progress in a level.

In general, the benefits of our approach to measuring problem solving in an engaging game stem from our use of evidence-centered design, which gives a framework for creating valid assessments, and stealth assessment, which gives us the ability to deeply embed such assessments into complex learning environments like games. By embracing evidence-centered design and stealth assessment, other researchers can also create complex and engaging assessments that meet their specific needs.

Acknowledgements

We would like to thank our colleagues at GlassLab (glasslab.org) for supporting our work assessing problem solving in Plants vs. Zombies 2—specifically Jessica Lindl, Liz Kline, Michelle Riconscente, Ben Dapkiewicz, and Michael John. We would also like to thank Katarina Krkovic for her efforts in collecting and preparing the MicroDYN data.

References

Almond, R. G. (2010). I can name that Bayesian network in two matrixes! *International Journal of Approximate Reasoning, 51*, 167–178.
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., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. New York: Springer-Verlag.
Anderson, J. R. (1980). *Cognitive psychology and its implications*. New York, NY: Freeman.
Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.
Bransford, J., & Stein, B. S. (1984). *The IDEAL problem solver: A guide for improving thinking, learning, and creativity*. New York, NY: W. H. Freeman.
Buelow, M. T., Okdie, B. M., & Cooper, A. B. (2015). The influence of video games on executive functions in college students. *Computers in Human Behavior, 45*, 228–234.
Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*, 66–69. http://dx.doi.org/10.1126/science.1167311.
Gagné, R. M. (1959). Problem solving and thinking. *Annual Review of Psychology, 10*, 147–172.
Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist, 21*, 99–120.
Greiff, S., & Funke, J. (2009). Measuring complex problem solving: the MicroDYN approach. In the transition to computer-based assessment: new approaches to skills assessment and implications for large-scale testing. In F. Scheuermann, & J. Björnsson (Eds.), *Office for official publications of the European communities, Luxembourg, Luxembourg* (pp. 157–163).
Greiff, S., & Wüstenberg, S. (2014). Assessment with microworlds using MicroDYN: measurement invariance and latent mean comparisons. *European Journal of Psychological Assessment, 30*(4), 304–314. http://dx.doi.org/10.1027/1015-5759/a000194.
Greiff, S., Wüstenberg, S., Csapo, B., Demetriou, A., Hautamäki, J., Graesser, A. C., et al. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13*, 74–83.
Hart Research Associates. (2015). *Falling short? College learning and career success*. Retrieved from: https://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf.
Hou, H. T., & Li, M. C. (2014). Evaluating multiple aspects of a digital educational problem-solving-based adventure game. *Computers in Human Behavior, 30*, 29–38.
Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63–85.
Jonassen, D. (2003). Using cognitive tools to represent problems. *Journal of Research on Technology in Education, 35*(3), 362–381.
Jonassen, D. H. (1997). Instructional design models far well-structured and ill-structured problem solving learning outcomes. *Educational Technology Research and Development, 45*(1), 65–94.
Jonassen, D. H., Marra, R., & 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.
Kiili, K. (2007). Foundation of problem-based gaming. *British Journal of Educational Technology, 38*(3), 394–404.
Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. Handbook of educational psychology. In D. C. Berliner, & R. C. Calfee (Eds.), *Macmillan library reference, New York, NY* (pp. 47–62).
Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander, & P. H. Winnie (Eds.), *Handbook of educational psychology* (pp. 287–303). Hillsdale, NJ: Lawrence Erlbaum.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.
Newell, A., & Shaw, J. C. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*(3), 151–166.
OECD. (2014). *PISA 2012 Results: Creative problem Solving: Students' skills in tackling real-life problems* (Vol. 5). Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264208070-en.
Partnership for the 21st Century (2012). http://www.p21.org.
Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton, NJ: Princeton University Press.
Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology, 19*(1), 137–150.
Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology, 41*, 1–48.
Ruscio, A. M., & Amabile, T. M. (1999). Effects of instructional style on problem-solving creativity. *Creativity Research Journal., 12*, 251–266.
Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences., 24*, 42–52.
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., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler,

D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 43–58). New York, NY: Springer.
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. (2016). Advances in the science of assessment. *Educational Assessment, 21*(1), 1–27.
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., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: flow and grow. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.
Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: the effects of portal 2 and luminosity on cognitive and noncognitive skills. *Computers & Education, 80*, 58–67.
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*, 423–430.
Shute, V.J., & Wang, L., Assessing and supporting hard-to-measure constructs. In A. Rupp, & J. Leighton (Eds.), *Handbook of cognition and assessment*. New York, NY: Springer (in press)
Taylor, L., & Parsons, J. (2011). Improving student engagement. *Current Issues in Education, 14*(1). Retrieved from http://cie.asu.edu/.
Van Eck, R. (2006). Building intelligent learning games. In Games and simulations in online learning: research & development frameworks. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Idea Group, Hershey, PA*.
Wang, L., Shute, V. J., & Moore, G. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer Mediated Simulations, 74*(4), 66–87.
Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving — more than reasoning? *Intelligence, 40*, 1–14.