Chapter 4
Assessment and Adaptation in Games

Valerie Shute, Fengfeng Ke, and Lubin Wang

Abstract Digital games are very popular in modern culture. We have been examining ways to leverage these engaging environments to assess and support important student competencies, especially those that are not optimally measured by traditional assessment formats. In this chapter, we describe a particular approach for assessing and supporting student learning in game environments—stealth assessment—that entails unobtrusively embedding assessments directly and invisibly into the gaming environment. Results of the assessment can be used for adaptation in the form of scaffolding, hints, and providing appropriately challenging levels. We delineate the main steps of game-based stealth assessment and illustrate the implementation of these steps via two cases. The first case focuses on developing stealth assessment for problem-solving skills in an existing game. The second case describes the integration of game and assessment design throughout game development, and the assessment and support of mathematical knowledge and skills. Both cases illustrate the applicability of data-driven, performance-based assessment in an interactive game as the basis for adaptation and for use in formal and informal contexts.

Keywords Stealth assessment • Adaptation • Bayesian networks

4.1 Introduction

According to "*2015 Essential Facts About the Computer and Video Game Industry*" published by Entertainment Software Association, over 150 million Americans play video games and 42 % play regularly for at least 3 h per week. The popularity of video games has drawn researchers' attention in the exploration of the possibility of using video games to enhance knowledge, skills, and other personal attributes. The idea of using games for serious purposes other than entertainment is called game-based learning. Advocates of game-based learning argue that well-designed

V. Shute (✉) • F. Ke • L. Wang
∎, ∎, ∎

P. Wouters, H. van Oostendorp (eds.), *Techniques to Improve the Effectiveness of Serious Games*, Advances in Game-Based Learning,
DOI 10.1007/978-3-319-39298-1_4

video games represent solid learning principles such as providing ongoing feedback, interactivity, meaningful and engaging contexts, and adaptive challenges within the zone of proximal development (Bransford, Brown, & Cocking, 2000; Gee, 2003; Shute, 2008; Vygotsky, 1978). A fair amount of research shows that game-based learning is at least as effective as nongame conditions, such as classroom contexts (e.g., Barab, Gresalfi, & Ingram-Goble, 2010; Clark, Tanner-Smith, & Killingsworth, 2014; Sitzmann, 2011; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013).

Researchers are also beginning to realize that games can serve as effective assessments (e.g., DiCerbo & Behrens, 2012; Shute, Leighton, Jang, & Chu, 2016; Shute & Ventura, 2013). That is, while players interact with the game environment, the game engine monitors and collects information about players' performances and provides feedback to players in the form of in-game scores or the avatar's progress in the game. This is basically the same as what educational assessment does, i.e., making inferences about students' knowledge and skills by observing what students say, do, and produce in a given context (Mislevy, Steinberg, & Almond, 2003). In addition, when game-based assessment is designed following a principled assessment design framework such as evidence-centered design (ECD; Mislevy et al., 2003) or cognitive design system (CDS; Embretson, 1998), the assessment is likely to have high validity and reliability.

Game-based assessment is essentially performance-based assessment. Performance-based assessment refers to tasks that require students to demonstrate their knowledge and skills by working through a task (Flynn, 2008; Madaus & O'Dwyer, 1999). Rather than a simple test of one's ability to recall or recognize information, or supply self-reported information, performance-based assessment provides students with the opportunity to show their understanding and apply knowledge in meaningful settings (Stecher, 2010). Scholars generally support the use of performance-based assessment to measure and support twenty-first-century skills (e.g., problem solving, creativity, collaboration; Partnership for the 21st Century 2015) over conventional types of assessment such as multiple-choice questions or filling in the blanks (see Shute et al., in press). However, there are a few challenges associated with the design and implementation of performance-based assessments. Some of the more difficult challenges include: (a) designing contexts that will fully elicit the competencies to be measured, (b) modeling the multidimensionality of constructs to be measured, (c) ensuring the validity and reliability (consistency) of the tasks, (d) providing appropriate feedback that is customized to each individual situation, (e) automating the scoring of the various tasks, (f) accumulating the evidence across all task performances, and (g) reducing the development costs of performance-based assessments compared to traditional tests. Our premise in this chapter is that stealth assessment (see Shute, 2011) coupled with ECD provides a viable solution to these challenges.

In addition to serving as assessment vehicles, games can help to support learning and motivation. That is, people who want to excel at something spend countless hours making intellectual effort and practicing their craft. But practice can be boring and frustrating, causing some learners to abandon their practice and, hence, learning.

This is where the principles of game design come in—good games can provide an engaging and authentic environment designed to keep practice meaningful and personally relevant. With simulated visualization, authentic problem solving, and instant feedback, computer games can afford a realistic framework for experimentation and situated understanding, and thus act as rich primers for active, motivated learning (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; Squire, 2006). Another key feature of well-designed games that can enhance learning and motivation is adaptivity related to providing appropriate and adaptive levels of challenge (see Fullerton, 2014). Gee (2003) has argued that the secret of a good game is not its 3D graphics and other bells and whistles, but its underlying architecture in which each level dances around the outer limits of the player's abilities, seeking at every point to be hard enough to be just doable. Similarly, psychologists (e.g., Vygotsky, 1987) have long argued that the best instruction hovers at the boundary of a student's competence. Flow is another name for this phenomenon. It is a construct first proposed by Csikszentmihalyi (1990, 1997) to describe an optimal experiential state that involves complete immersion in an activity and a deep sense of enjoyment. Flow represents full engagement, which is crucial for deep learning. The essential components of flow include clear and unambiguous goals, challenging yet achievable levels of difficulty, and immediate feedback (Cowley, Charles, Black, & Hickey, 2008; Csikszentmihalyi, 1997). In the game design context, flow theory states that if the player finds a level too difficult, he/she will become frustrated. However, if, as the player continues playing, his/her abilities improve while the challenge level stays the same, he/she will become bored. Therefore, to facilitate a flow state, challenge and ability must be carefully balanced to accomplish this type of adaptivity.

In this chapter, we first review the theoretical foundations of ECD and stealth assessment. In the second section, we discuss how stealth assessment works. After the discussion, we demonstrate the process of creating stealth assessment using ECD via two examples—one past and one current research project—that apply the approach. We then conclude this paper with a brief discussion on implications for future research.

4.2 Literature Review

4.2.1 Evidence-Centered Design

Evidence-centered design (Mislevy et al., 2003) provides a framework for designing and implementing assessments that support arguments about personal competencies via an evidence chain that connects the arguments with task performance. ECD consists of conceptual and computational models that work together. The three major models include the competency model, the evidence model, and the task model.

The *competency model* outlines in a structured fashion the beliefs about personal knowledge, skills, or other learner attributes. The competency model can host unidimensional constructs and, importantly, multidimensional constructs (e.g., problem solving, leadership, and communication skills) as well. The beliefs about learners' competencies in the competency model are updated as new evidence supplied by the evidence model comes in. When competency model variables are instantiated with individual student data, the competency model is often referred to as the student model.

The *task model* identifies the features of selected tasks for learners that will provide evidence about their target competencies. The main function of the task model is to provide observable evidence ,about the unobservable competencies, which is realized via the evidence model.

The *evidence model* serves as the bridge between the competency model and the task model. It transmits evidence elicited by tasks specified by the task model to the competency model by connecting the evidence model variables and competency model variables statistically. Basically, the evidence model contains two parts: (a) evidence rules or rubrics that convert the work products created during the interactions between the learner and the tasks to observable variables that can be scored in the form of "correct/incorrect" or graded responses; and (b) a statistical model that defines the relationships among observable variables and competency model variables, and then aggregates and updates scores across different tasks. The statistical model may be in the form of probabilities based on Bayes theorem or they may be simple cut scores.

4.2.2 Stealth Assessment

Stealth assessment, a specialized implementation of ECD, is a method of embedding assessment into a learning environment (e.g., video games) so that it becomes invisible to the learners being assessed (Shute, 2011). We advocate the use of stealth assessment because of its many advantages. As we mentioned at the beginning of the chapter, there are a number of challenges related to performance-based assessment, but stealth assessment addresses each challenge. Because it is designed to be unobtrusive, stealth assessment frees students from test anxiety commonly associated with traditional tests and thus improves the reliability and validity of the assessment (e.g., DiCerbo & Behrens, 2012; Shute, Hansen, & Almond, 2008). Second, stealth assessment is designed to extract ongoing evidence and update beliefs about students' abilities as they interact with the tasks. This allows assessors to diagnose students' performance and provide timely feedback. As a result, interacting with the learning or gaming environment can support the development of students' competencies as they are being assessed. Third, when stealth assessment is designed following ECD, this allows for the collection of sufficient data about students' target competencies at a fine grain size providing more information about a student's ability compared with conventional types of assessment like multiple-choice formats. Fourth, when stealth assessment is embedded within a well-designed video game, students are fully engaged in the experience, which is conducive to the extraction of true knowledge and skills. Fifth, because scoring in stealth assessment is automated, teachers do not need to spend valuable time calculating scores and grades. Finally, stealth assessment models, once developed and validated, can be reused in other learning or gaming environments with only some adjustments to the particular game indicators.

Recently, we have been creating and testing stealth assessments of various competencies within video games. For instance, we developed and embedded three stealth assessments (running concurrently) of qualitative physics understanding (Shute, Ventura, & Kim, 2013), persistence (Ventura, Shute, & Small, 2014; Ventura, Shute, & Zhao, 2012), and creativity (Kim & Shute, in press) in a homemade game called *Physics Playground*, formerly called *Newton's Playground* (see Shute & Ventura, 2013). We created and tested stealth assessments of problem solving and spatial skills for the commercial game *Portal 2* (Shute, Ventura, & Ke, 2015; Shute & Wang, in press). Additionally, we created stealth assessment of causal reasoning in the *World of Goo* (Shute & Kim, 2011) and systems thinking in *Taiga Park* (Shute, Masduki, & Donmez, 2010). From these experiences, we have derived some general steps related to the design and development of stealth assessment, shown in the 9-step approach listed as follows. In the following section, we illustrate how we implemented these steps using two recent research projects.

1. Develop competency model (CM) of targeted knowledge, skills, or other attributes based on full literature and expert reviews
2. Determine which game (or learning environment) the stealth assessment will be embedded into
3. Delineate a full list of relevant gameplay actions/indicators that serve as evidence to inform CM and its facets
4. Create new tasks in the game, if necessary (Task model, TM)
5. Create Q-matrix to link actions/indicators to relevant facets of target competencies
6. Determine how to score indicators using classification into discrete categories (e.g., yes/no, very good/good/ok/poor relative to quality of the actions). This becomes the "scoring rules" part of the evidence model (EM)
7. Establish statistical relationships between each indicator and associated levels of CM variables (EM)
8. Pilot test Bayesian Networks (BNs) and modify parameters
9. Validate the stealth assessment with external measures

4.2.3 Adaptation

The next logical step—which is currently under development—involves using the current information about a player's competency states to provide adaptive learning support (e.g., targeted formative feedback, progressively harder levels relative to the player's abilities, and so on). The adaptive difficulty features in a video game may potentially increase motivation and enhance learning by providing the right level of challenge (i.e., tasks that are neither too easy nor too difficult). Such optimal levels of challenge ensure that the learner is kept in the zone of proximal development (ZPD). Within ZPD, learning activities are just beyond the learner's ability but can be achieved with guidance (Vygotsky, 1978). The guidance is sometimes referred to as instructional scaffolding. Some examples of such scaffolding include targeted formative feedback and hints to help learners proceed in the task. Studies show that scaffolded learning activities lead to better learning outcomes compared with activities without scaffolds (e.g., Chang, Sung, & Chen, 2001; Murphy & Messer, 2000). In addition, when tasks are too complicated for a learner, he or she may encounter cognitive overload that exceeds the capacity of their working memory and thus undermines learning. On the other hand, if the tasks are too easy, the learner may feel bored and disengaged, which also negatively affects learning. Therefore, it is important and beneficial to adjust the difficulty of tasks to the competencies of the individual and provide appropriate learning scaffolds.

There are two main approaches to produce adapted content in video games—offline and online adaptivity (Lopes & Bidarra, 2011). For offline adaptivity, content is adjusted after gathering sufficient information about the learner before he or she starts playing the game. For online adaptivity (or dynamic adaptivity; see van Oostendorp, van der Spek, & Linssen, 2014), the content is adjusted based on learner's performance, in real time. We recommend the second approach because the assessment of the learner's competency will be more accurate when he or she is actually performing the task.

Some common ways to gather information about the learner during gameplay include the use of infrared camera or emotion detection software, and stealth assessment. One issue with infrared camera or emotion detection software is that different people may experience different levels of stress when they are under pressure. Thus, it is difficult to choose the right task based on the stress level. Alternatively, stealth assessment gathers data unobtrusively based on performance in the game and is free from such bias.

To determine the sequence of tasks in video games, researchers have attempted to set an agreed-upon threshold value (e.g., level up after three consecutive successes; see Sampayo-Vargas, Cope, He, & Byrne, 2013). Some have calculated the expected weight of evidence to pick tasks that will maximize the information about a player (Shute et al., 2008). Due to the relatively high cost of developing adaptive educational games, few researchers have attempted to investigate the effects of adaptive video games on learning. However, existing evidence shows that such methods are promising. For example, van Oostendorp et al. (2014) compared the effects of an adaptive version of a game focusing on triage training against a version without adaptation. They reported that those who played the adaptive version of the game learned better than those in the control group.

4.3 Examples of Stealth Assessment

4.3.1 "Use Your Brainz" (UYB)

4.3.1.1 Competency Model Development and Game Selection (Steps 1 and 2)

In the UYB project, we developed a stealth assessment of problem-solving skills and embedded it within the modified version of the commercial game *Plants* vs. *Zombies 2* (the education version is called "Use your Brainz"). The project was a joint effort between our research team and GlassLab. PvZ 2 is a tower defense type of game. The goal is to protect the home base from the invasion of zombies by planting various defensive and offensive plants in the limited soil in front of the home base. We selected 43 game levels arranged by difficulty. Figure 4.1 shows an example of one of the levels in the game.

We chose the game PvZ 2 for two main reasons. First, the game provides a meaningful and engaging context where players are expected to acquire knowledge about the rules of the game and apply different resources in the game to solve intriguing problems. Second, GlassLab had access to the source code from EA—the publisher of PvZ 2—which enabled us to customize the log files.

Fig. 4.1 Screen capture of UYB gameplay on Level 9, World 1 (Ancient Egypt)

After we determined that we would like to model problem-solving skills, we reviewed the literature on how other researchers have conceptualized and operationalized problem solving. In addition to our extensive review of the literature on problem-solving skills, we also reviewed the Common Core State Standards (CCSS) related to problem solving. We came up with a four-facet competency model (CM), which included: (a) understanding givens and constraints, (b) planning a solution pathway, (c) using tools effectively/efficiently when implementing solutions, and (d) monitoring and evaluating progress.

4.3.1.2 Identifying Gameplay Indicators (Steps 3 and 4)

Our next task entailed identifying specific in-game behaviors that would serve as valid evidence and thus inform the status of the four-facet competency model. After playing the game repeatedly and watching expert solutions on YouTube, we delineated 32 observable indicators that were associated with the four facets. For example, sunflowers produce sun power, which is the sole source of power that players may use to grow plants. At the beginning of a level, typically there are no or very few sunflowers on the battlefield. To supply power to grow plants, players must plant sunflowers at the beginning of each level before zombies start to appear in waves. After brainstorming with the PvZ 2 experts on our research team, we decided that the scoring rule for this particular indicator was: "*If a player plants more than three sunflowers before the second wave of zombies arrives*, *the student understands the time and resource constraints*." Table 4.1 displays a sample of indicators for each of the four problem-solving facets. Overall, we included 7 indicators for "analyzing givens and constraints," 7 for "planning a solution pathway," 14 for "using tools effectively and efficiently," and 4 for "monitoring and evaluating progress." The list of indicators forms our task model and the scoring rules form a part of the evidence model.

Table 4.1 Examples of indicators for each problem-solving facet

Facet	Example indicators
Analyzing givens and 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/left, offensive plants in the middle, and defensive plants up front/right
	• Plants Twin Sunflowers or uses plant food on (Twin) Sunflowers in levels that require the production of X amount of sun
Using tools effectively and efficiently	• Uses plant food when there are >5 zombies in the yard or zombies are getting close to the house (within two 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

4.3.1.3 Q-Matrix Development and Scoring Rules (Steps 5 and 6)

We created a Q-matrix (Almond, 2010; Tatsuoka, 1990) laying out all of the indicators in rows and the four facets in the columns. We added a "1" in the crossed cell if the indicator was relevant to the facet and "0" if the facet did not apply to the indicator. We then went through each indicator and discussed how we could classify each indicator into discrete scoring categories such as "yes/no" or "very good/good/ok/poor." The overall scoring rules were based on a tally of relevant instances of observables. Using the aforementioned sunflower indicator, if a player successfully planted more than three sunflowers before the second wave of zombies arrived on the scene, the log file would automatically record the action and categorize it as a "yes" status of the indicator.

For another example, consider the facet "using tools effectively and efficiently." In Table 4.1, an example indicator is "uses plant food when there are >5 zombies in the yard or zombies are getting close to the house (within two squares)." Plant food in the game is a rare resource. Using one dose of plant food on any plant will substantially boost the effect of the plant—whether offensive or defensive—for a short period of time. This indicator would be scored if the player used plant food as a boost (a) when there were more than five zombies on the battlefield, or (b) when zombies were within two squares in front of the house (where the overarching goal of each level is to protect the house from zombies). Since a single instance of this "using plant food" action may be performed by chance, the completion status of the indicator was categorized into four levels. That is, the game engine checks on the ratio of the indicator, which is "the number of times that plant food was used when >5 zombies in the yard or within two squares in front of the house, divided by the total number of times that plant food was used in the level." Then the game engine maps the value of the ratio onto one of the four states of the indicator where in this case, higher means better. If the value is within [0, 0.25], it corresponds to the status of "poor" performance on the indicator; if the value falls within [0.26, 0.5], it corresponds to the "ok" status; if the value falls within [0.51, 0.75], it corresponds to the "good" status, and if the ratio falls within [0.76, 1], it is categorized as "very good."

4.3.1.4 Establishing Statistical Relationships Between Indicators and CM Variables (Step 7)

Once we categorized all indicators into various states, we needed to establish statistical relationships between each indicator and the associated levels of the CM variables. We used Bayesian networks (BNs) to accumulate incoming data from gameplay and update beliefs in the CM. The relationship between each indicator and its associated CM variable was expressed within conditional probability tables stored in each Bayes net. We created a total of 43 Bayes nets for this project, one for each level. We used separate BNs because many indicators do not apply in every level and computations would be more efficient for simpler networks. The statistical relationships carried in the Bayes nets and the scoring rules described in the last section formed the evidence model.

Table 4.2 Conditional probability table for indicator #8 "plant >3 sunflowers before the second wave of zombies" in Level 9

Analyzing givens and constraints	Yes	No
High	.82	.18
Medium	.73	.27
Low	.63	.37

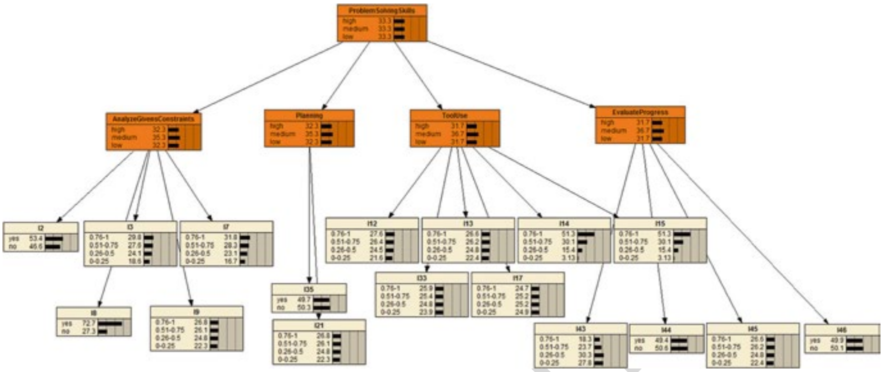

 Fig. 4.2 Bayes network of level 9 in UYB, prior probabilities

Table 4.2 shows the conditional probability table we created for indicator #8, "Plants >3 Sunflowers before the second wave of zombies arrives" (associated with the facet "analyzing givens and constraints") in Level 9. Because the game is linear (i.e., you need to solve the current level before moving to the next level), by the time a player gets to Level 9, she has had experience playing previous levels, thus should be quite familiar with the constraint of planting sunflowers at this point. Consequently, this indicator should be relatively easy to accomplish (i.e., the probabilities to fail the indicator were low despite one's ability to analyze givens and constraints). Even those who are low on the facet still have a probability of .63 of accomplishing this indicator. When evidence about a student's observed results on indicator #8 arrives from the log file, the estimates on his ability to analyze givens and constraints will be updated based on Bayes theorem. We configured the distributions of conditional probabilities for each row in Table 4.2 based on Samejima's graded response model, which includes the item response theory parameters of discrimination and difficulty (see Almond, 2010; Almond et al., 2001; Almond, Mislevy, Steinberg, Williamson, & Yan, 2015). In this case, the difficulty was set at –2 (very easy) and the discrimination value was 0.3 (i.e., may not separate students with high versus low abilities well).

As a player interacts with the game, incoming evidence about the player's status on certain indicators updates the estimates about relevant facets. The evidence then propagates through the whole network and thus estimates related to student problem-solving skills are updated. The Bayes nets keep accumulating data from the indicators and updating probability distributions of nodes in the network. For example, Fig. 4.2 displays a full Bayes net of Level 9 prior probabilities (see Fig. 4.1 for an illustration of the level). Shaded nodes toward the top are the competency

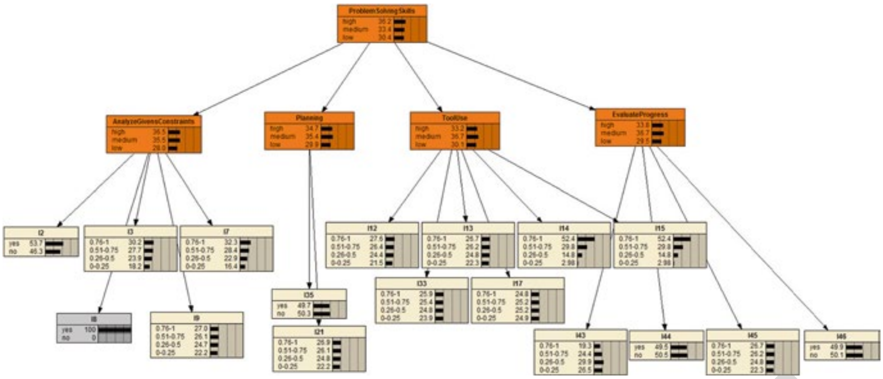

Fig. 4.3 Evidence of the completion of indicator #8

variables, while the beige nodes toward the bottom represent all relevant indicators. We used the program Netica (by Norsys Software Corporation) to construct and compile the network.

For instance, if a player successfully completed indicator #8 in Level 9 (i.e., planting sufficient sunflowers prior to a wave of incoming zombies), the log file records the action, informs the network of the new evidence, and the data are propagated throughout the network (see Fig. 4.3). As shown, the updated probability distribution of the player's level of "analyzing givens and constraints" is: Pr (analyzing givens and constraints | high) = .365, Pr (analyzing givens and constraints | med) = .355, Pr (analyzing givens and constraints | low) = .280. The estimates for the player's overall problem-solving skill are Pr (problem solving | high) = .362, Pr (problem solving | med) = .334, Pr (problem solving | low) = .304. Because there is no clear modal state for the problem-solving skills node (i.e., the difference between high and medium states is just .028), this suggests that more data are needed.

Alternatively, suppose the player fails to accomplish the indicator by the second wave of zombies. In this case, the log file would record the failure, inform the BN of the evidence, and update with new probability distributions for each node (Fig. 4.4). The current probability distribution of the player's level of "analyzing givens and constraints" is Pr (analyzing givens and constraints | high) = .213, Pr (analyzing givens and constraints | med) = .349, Pr (analyzing givens and constraints | low) = .438. The estimates for the player's overall problem solving skill are Pr (problem solving | high) = .258, Pr (problem solving | med) = .331, Pr (problem solving | low) = .411. This shows that the student is likely to be low in relation to problem-solving skills.

4.3.1.5 Pilot Testing Bayes Nets (Step 8)

Our game experts and psychometricians produced the initial prior probabilities of each node in each network collaboratively. We hypothesized that students would have an equal likelihood of being "high," "medium," or "low" on problem solving

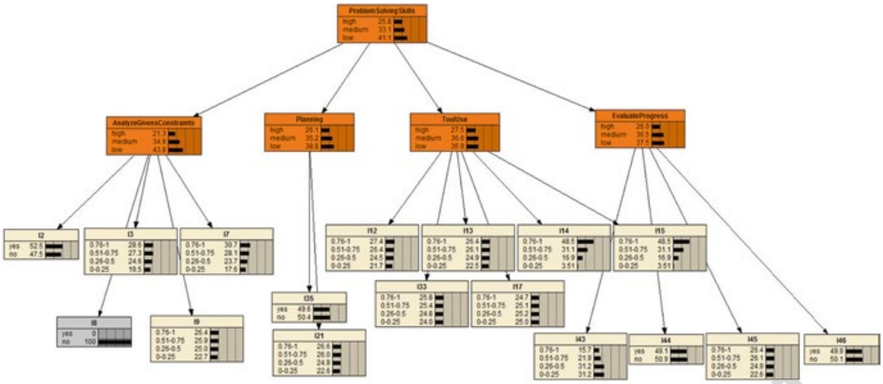

Fig. 4.4 Evidence of failure to complete indicator #8

and the probability of being "high," "medium," or "low" for each facet would be normally distributed. As more evidence enters the network, the estimates become more accurate and tend to reflect each player's true status on the competency. After developing the BNs and integrating them into the game code, we were able to acquire real-time estimates of players' competency levels across the main node (problem-solving skill) and its constituent facets. We acknowledge that any initial probabilities may be subject to bias or inaccurate judgment. Therefore, we ran a pilot test and used the ensuing pilot data to adjust parameters of the Bayes nets accordingly.

4.3.1.6 Validating Stealth Assessment (Step 9)

The final step in our list of stealth assessment processes is the validation of the stealth assessment against external measures. For the UYB project, we employed two external measures: *Raven's Progressive Matrices* (Raven, 1941, 2000) and *MicroDYN* (Wustenberg, Greiff, & Funke, 2012). Raven's is a test that examines subjects' ability to reason based on given information. MicroDYN presents to subjects a simulation system where subjects are expected to acquire and apply information. For a thorough overview on MicroDYN, see Schweizer, Wüstenberg, and Greiff (2013) and Wustenberg, Greiff, and Funke (2012).

We recruited 55 7th grade students from a middle school in suburban Illinois. Students played UYB for 3 h (1 h per day across three consecutive days) and completed the external measures on the fourth day. Among the 55 participants, one student's gameplay data was missing, five students did not take the Raven's test, and two students did not complete the MicroDYN test. After we removed the missing data, we had complete data from 47 students (20 male, 27 female).

Results show that our game-based stealth assessment of problem-solving skills is significantly correlated with both Raven's ($r = .40$, $p < .01$) and MicroDYN ($r = .41$, $p < .01$), which established the construct validity of our stealth assessment. We are also refining our Bayes nets based on data collected. These test results need to be verified with an even larger sample.

This example demonstrates step by step how we modeled problem-solving skills and created and implemented stealth assessment of the skill in the context of a modified commercial game. Specifically, we created our competency model of problem-solving skills based on the literature, identified relevant indicators from gameplay that could provide evidence of players' levels on the competency model variables, crafted scoring rules of each indicator, and connected the indicators statistically with competency model variables. We then modified the Bayes networks by collecting and analyzing data collected from a pilot study. Then, we selected well-established external measures and validated the stealth assessment in a validation study. Reasonable next steps would entail developing tools to help educators gain access to the results of the assessment easily (e.g., via a dashboard displaying and explaining important results). With that information, educators could effectively and efficiently support the growth of problem-solving skill, at the facet level.

4.3.2 "Earthquake Rebuild" (E-Rebuild)

As discussed in the preceding example with UYB, the stealth assessment was designed and implemented as a post-hoc practice because the game had already been designed. In a current design-based project (called Earthquake Rebuild), we have been designing evidenced-centered stealth assessment during the entire course of game design. Earthquake Rebuild (E-Rebuild) acts as both a testbed and sandbox for generating, testing, and refining the focus design conjectures on game-design-associated, stealth assessment and support of learning.

Developed using Unity 3D, the overall goal of E-Rebuild is to rebuild an earthquake-damaged space to fulfill diverse design parameters and needs. The intermediate game goal involves completing the design quest(s) in each game episode to gain new tools, construction materials, and credits. A learner in E-Rebuild performs two modes of play: (a) third-person construction mode, and (b) first-person adventure mode. In the third-person construction mode, a learner performs construct site survey and measurement and maneuver (e.g., cut/scale, rotate, and stack up) construction items to build the targeted structure. In the adventure mode, a learner navigates the virtual world, collects or trade construction items, and assigns space (to residents, for example).

The process of interweaving game and assessment design in E-Rebuild included four core design sectors: (1) developing competency models and selecting game mechanics that necessitate the performance of the focus competency, (2) designing game task templates and contextual scenarios along with the Q-matrix, (3) designing the game log file based on the Q-matrix, and (4) designing the in-game support as both live input for data-driven assessment and adaptive feedback. These design sectors are interacting and interdependent with each other.

4.3.2.1 Competency Model and Game Mechanics Development

In E-Rebuild, an interdisciplinary team of math educator, mathematician, and assessment experts codeveloped a competency model for each focal math topic. These competency models are aligned with the Common Core State Standards (CCSS) for mathematical practice in grades 6–8. The game design team then designed and selected game mechanics that would best serve the competency models. Specifically, game actions were the core constituent of game mechanics and the basic behavioral unit to be tracked during gameplay. Consequently, game actions became the driving element, defining the degree of learning integration and assessment in the game. The team focused on designing game actions or indicators that would *necessitate*, *not just allow*, the performance of focus knowledge and skills (e.g., ratio and proportional reasoning). By experimenting with all proposed architectural design actions via iterative expert review and user testing at the initial paper prototyping stage, the design team decided on the game actions that best operationalized the practice of math knowledge, which include (material) trading, building, and (resource) allocation. Furthermore, comparative analyses with different versions of the game prototype in a one-year case study indicated that an intermediary yet noninterruptive user input (e.g., entering a specific number), in comparison with an intuitive user input (e.g., clicking or dragging a button or meter to adjust a numerical value), effectively necessitates the practice of the targeted mathematical knowledge. For example, the trading interface (see Fig. 4.5) requires the player to enter the quantity of a building item to be ordered, calculate the total amount/cost (based on the unit rate), and enter the numerical value. Similarly, the scaling tool prompts the player to specify the numerical value for the scaling factor to scale down a 3D construction item along the chosen local axis of the item (x, y, z, or all).

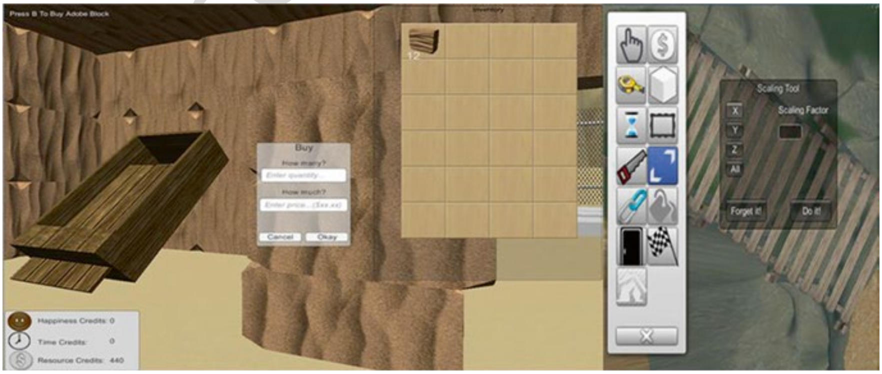

Fig. 4.5 Intermediary user input example—the trading interface and the scaling tool for the building action

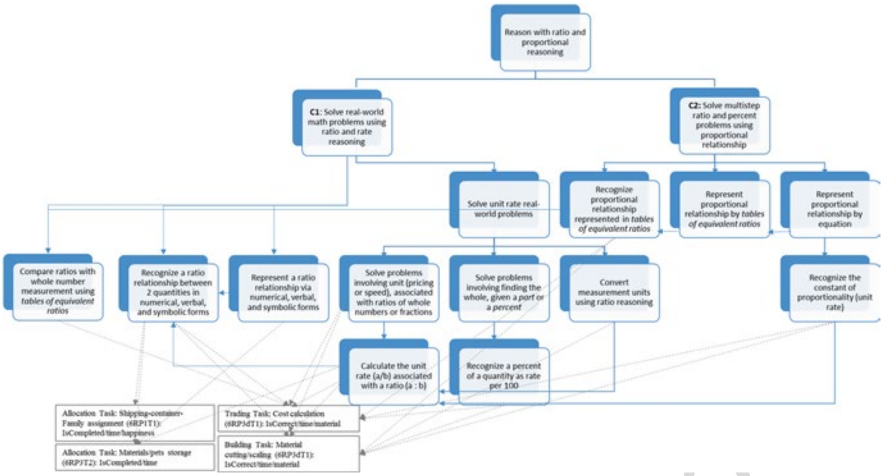

Fig. 4.6 A design document depicting a competency model along with the design of game task templates. *Note*: The four *black boxes* at the bottom represent examples of game tasks designed to extract the subcompetencies, which are depicted in the *blue boxes* in a hierarchical structure. *Solid lines* indicate the relationships among competencies and subcompetencies to be captured/assessed, and *dotted lines* link the gaming tasks and the competencies to be assessed.

4.3.2.2 Designing Task Templates to Substantiate the Competency Model and Q-Matrix

In E-Rebuild, the game task development was confined by the math competency models. Specifically, the competency model has driven the development of a cluster of game task templates and the selection of the tasks' parameters and content scope (as depicted in Fig. 4.6). For instance, an exemplary allocation task (e.g., assigning families into a multiroom shelter structure, with the ratio of an adult's living space need to a child's need being 2 to 1) was designed to extract math performance of subcompetencies (e.g., C1) of "ratio and proportional reasoning." The Q-matrix development (Fig. 4.7) then helped the design team gauge and track which facets of the math competency a specific gameplay action inform, and whether each facet of a math competency is practiced/assessed by different clusters of tasks. Accordingly, existing task templates could be refined or removed, and new task templates might be developed.

The Q-matrix also helped the team to gauge the discrimination and difficulty qualities of different tasks and hence assisted the selection and sequencing of tasks within/across game episodes. Finally, a variety of architecture-themed scenarios (e.g., building shelters with shipping containers or building a structure to meet the needs of multiple families) would contextualize different clusters of game tasks and inform the development of the task narrative. These aforementioned design processes occurred concurrently and helped to make the game-task design and the evidence model development a coherent process.

Task Name	ObsName	Reason with ratio and proportional reasoning								
		Compare ratios with whole number measurement using tables of equivalent ratios	Recognize a ratio relationship between 2 quantities in numerical form	Recognize a ratio relationship between 2 quantities in verbal form	Recognize a ratio relationship between 2 quantities in symbolic form	Represent a ratio relationship via numerical form	Represent a ratio relationship via verbal form	Represent a ratio relationship via symbolic form	Calculate the unit rate (a/b) associated with a ratio (a : b)	Recognize a percent of a quantity as rate per 100
Allocation Task	timeToCompletion	0	1	1	1	1	0	1	1	0
	Material Credit	0	0	0	0	0	0	0	1	0
	scratchpad editing(math related)	0	0	0	0	1	0	0	1	0
	assignment operation	0	0	0	1	0	0	1	1	0
Trading Task	# of trades	1	1	1	0	1	0	0	1	1
	scratchpad editing(math related)	0	0	0	0	1	0	0	1	0
	percentage lost in trade avg	1	1	1	0	1	0	0	1	1
	cut (for resourcing)	0	0	0	0	0	0	0	0	0
	scale (for resourcing)	0	0	0	0	0	0	0	0	0
Building Task	structure size	0	0	1	0	0	0	1	0	1
	structure location	0	0	0	0	0	0	0	0	0
	structure direction	0	0	0	0	0	0	0	0	0
	# copy/paste failed	0	0	0	0	0	0	0	0	0
	scratchpad editing(math related)	0	0	0	0	1	0	0	1	0
	ruler record	0	0	1	0	0	0	1	0	0
Game Task	timeToCompletion	1	1	1	1	1	0	1	1	1
	Material Credit	1	1	1	1	1	0	1	1	1
	Happiness Credit	0	1	1	1	1	0	1	1	1

Fig. 4.7 Part of the Q-matrix for E-Rebuild. *Note*: Facets of the focus competency are listed in columns and the indicators are listed in rows.

4.3.2.3 Designing Game Log File Along with Q-Matrix for Bayesian Network Construction

During the course of E-Rebuild design, we designed, tested, and refined the game log file along with the Q-matrix so that the game objects, salient object features, play actions, and action-performing statuses tracked in the game log will assist the generation and update of conditional probability tables (CPTs) for all indicators in the Bayes net being constructed. In E-Rebuild, the creation of CPTs for indicators and hence the Bayesian Network construction were initially driven by the logged gameplay data of 42 middle school students and 6 game/content experts in a pilot study. The CPTs and the preliminary networks generated were then reviewed and refined by the content/assessment experts and game designers. Game logs and indicators were also refined based on the pilot-testing results. For the next phase, the refined CPTs and Bayesian networks will be further tested and updated by the gameplay data to be collected from a larger group of target users, and then validated by external math knowledge tests in a future evaluation study.

4.3.2.4 *In-Game Support as Both Input and Output of Data-Driven Learning Assessment*

In E-Rebuild, we have designed in-game cognitive support (scaffolding) as an expandable/collapsible help panel and a scratch pad. The scratch pad includes an internal calculator and enables/records participants' typing of numerical calculation steps. The help panel (Fig. 4.8) contains interactive probes to facilitate active math problem representation rather than passively presenting the information. When

Fig. 4.8 Interactive learning probes

interacting with those probes, a player has to enter numbers or maneuver dynamic icons, with all interactions logged. The two support features thus work as another dynamic data source for game-based stealth assessment. In addition, we are still designing the dynamic-help mechanism that will use the values extracted from the logged gameplay performance variables (e.g., timeToCompletion, materialCredit, assignmentScore, usedScratchpad, helpInput) to inform the content and presentation of task-specific learner feedback in the Help panel. Based on the dynamically updated game task performance of the player, the game-based assessment mechanism will inform on task-relevant math competency (e.g., below 50 % in a specific competency). Accordingly, the help menu will be displayed automatically and a math-competency-related subsection of the problem-solving probes will be expanded. The interactive probes may be presented in iconic (pictorial) and/or symbolic (numerical formula) formats, pending on the player's choice.

4.4 Discussion and Implications

In this chapter, we have introduced the core steps of game-based stealth assessment of learning and illustrated the implementation of these steps via two cases. The first case focuses on developing an assessment mechanism for an existing game and the assessment of an important domain-general skill (i.e., problem solving). The second case highlights the integration of learning task and assessment design throughout the game development process and the assessment of domain-specific (mathematical) practice and learning. Both cases illustrate the applicability of data-driven, performance-based assessment in an interactive learning setting, for either formal or informal learning.

Several design challenges of in-game learning assessment should be considered. First, the development of the underlying competency model is critical for the (construct) validity of the game-based stealth assessment. The latent and observed competency variables, as well as the scope of the focal competency are usually confined by the literature base, the content expertise/background of the project team, and an external evaluation purpose or standard (e.g., Common Core State Standards in E-Rebuild). The competency model variables and scope are also moderated by the targeted learners and levels of learning outcomes. Hence the effort contributed to developing and validating the competency model is critical, and a developed competency model for assessment should be reviewed and refined for each implementation setting. Second, although the development of a global, overarching Bayesian network is desirable, creating individual Bayes nets for each game episode may be necessary to enhance the efficiency in data accumulation and nodes updating in the Bayesian net. Third, the creation of conditional probability tables for the initial construction of the Bayes net(s) should be driven by both expert opinion and in-field gameplay data.

In the first game (Use Your Brain), expert opinions drove the initial CPT development, which were then enhanced by in-field data validation. In E-Rebuild, CPTs were generated (learned) from the in-field data and then reviewed/refined by experts. Future research can experiment with the two methods in CPT generation and further investigate the potential differences in the two methods on learning and validating the Bayesian network. Finally, in both projects we are presently developing and testing various adaptive learning support mechanisms. The dynamically updated learning assessment in E-Rebuild will be used to drive the timing (e.g., at the end of a game action, a task, or a game level), topic (e.g., on a task-specific math concept or a calculation procedure), and the presentation format (e.g., iconic or symbolic, informative hint or interactive probe) of the learning scaffolds for game-based learning. A critical design consideration for assessment-based, dynamic learner support is the timing and extent of live data accumulation for adaptive support presentation. In E-Rebuild, we have used game level and game episode (i.e., an episode includes multiple game levels) as two hierarchical units for data accumulation and learning support presentation. Specifically, performance data will be fed into the Bayesian network at the end of each game level and each game episode. Correspondingly, the learner profile will be updated at these points, and then the relevant learner supports (e.g., probes and feedback) can be presented as both cut-screen in between game levels/episodes, and updated content in the Help panel.

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.

Almond, R. G., DiBello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., et al. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola & T. Richardson (Eds.), *Artificial intelligence and statistics 2001* (pp. 137–143). San Francisco, CA: Morgan Kaufmann.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Williamson, D. M., & Yan, D. (2015). *Bayesian networks in educational assessment*. New York: Springer.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play using games to position person, content, and context. *Educational Researcher, 39*(7), 525–536.

Barab, S. A., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development, 53*(1), 86–108.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). Washington: National Academies Press.

Chang, K. E., Sung, Y. T., & Chen, S. F. (2001). Learning through computer-based concept mapping with scaffolding aid. *Journal of Computer Assisted Learning, 17*, 21–33.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. (2014). *Digital games, design, and learning: A systematic review and meta-analysis*. Menlo Park, CA: SRI International.

Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment, 6*(2), 1–27.

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

Csikszentmihalyi, M. (1997). *Finding flow*. New York: Basic.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 300–396. doi:10.1037/1082-989X.3.3.380.

Entertainment Software Association. (2015). *2015 Essential facts about the computer and video game industry*. Retrieved from http://www.theesa.com/wp-content/uploads/2015/04/ESA-Essential-Facts-2015.pdf

Flynn, L. (2008). In praise of performance-based assessments. *Science and Children, 45*(8), 32–35.

Fullerton, T. (2014). *Game design workshop, 3rd edition: A playcentric approach to creating innovative games*. Boca Raton, FL: AK Peters/CRC Press.

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

Kim, Y. J., & Shute, V. J. (2015). Opportunities and challenges in assessing and supporting creativity in video games. In J. Kaufmann & G. Green (Eds.), *Research frontiers in creativity*. San Diego, CA: Academic.

Lopes, R., & Bidarra, R. (2011). Adaptivity challenges in games and simulations: A survey. *IEEE Transactions on Computational Intelligence and AI in Games, 3*(2), 85–99.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688–695.

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

Murphy, N., & Messer, D. (2000). Differential benefits from scaffolding and children working alone. *Educational Psychology, 20*(1), 17–31.

Partnership for the 21st Century. (2015). Retrieved from http://www.p21.org/storage/documents/P21_framework_0515.pdf

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.

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education, 69*, 452–462.

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. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

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., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education, 18*(4), 289–316.

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., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity 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. (in press). Assessing and supporting hard-to-measure constructs. In A. Rupp, & J. Leighton (Eds.), *Handbook of cognition and assessment*. New York, NY: Springer.

Sitzmann, T. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin, 137*, 421–442.

Squire, K. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19–29.

Stecher, B. (2010). *Performance assessment in an era of standard-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

van Oostendorp, H., van der Spek, E. D., & Linssen, J. (2014). Adapting the complexity level of a serious game to the proficiency of players. *EAI Endorsed Transactions on Serious Games, 1*(2), 8–15.

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 adaptive intelligent tutoring systems* (Learner modeling, Vol. 2, pp. 93–101). Orlando, FL: U.S. Army Research Laboratory.

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

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

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

Wouters, P. J. M., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*, 249–265.

Wustenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence, 40*, 1–14.