CHAPTER 20

STEALTH ASSESSMENT IN COMPUTER-BASED GAMES TO SUPPORT LEARNING

Valerie J. Shute
Florida State University

INTRODUCTION

This chapter is about stealth assessment: what it is, why it's needed, and how to accomplish it effectively. To make the ideas come alive, I provide two examples of stealth assessment in existing computer-based games. I end with my thoughts about challenges and next steps relating to this research stream.

The first time I formally used the term "stealth assessment" was in 2005, during an AERA symposium on diagnostic assessment. However, I had designed and employed stealth assessment about two decades prior to that, as part of a guided-discovery world called Smithtown (e.g., Shute & Glaser, 1990; Shute & Glaser, 1991; Shute, Glaser, & Raghavan, 1989). In Smithtown, students learned about principles of microeconomics (i.e., the laws of supply and demand) as they explored the simulated world, manipulated variables (e.g., the per capita income, population, price of coffee), tabulated and graphed data, and generated hypotheses about ensuing change(s)

Computer Games and Instruction, pages 503–523

to other variables based on their manipulations. The system used artificial intelligence methods to monitor and analyze student performance data relative to their scientific inquiry skills and provided feedback to students that supported these skills. The idea was that improving students' inquiry skills would subsequently improve their learning of the microeconomics content. A series of experiments supported the efficacy of this approach.

Now fast-forward to the present. Technologies (both hard and soft; see Shute & Zapata-Rivera, 2008), along with educational and psychological measurement approaches, have advanced a lot in the past couple of decades. We now can more accurately and efficiently diagnose student competencies at various levels during the course of learning. With regard to *low-level diagnoses* (i.e., at the problem or task level, addressing how the person handled a given problem), new technologies allow us to embed assessments into the learning process; extract ongoing, multifaceted information (evidence) from a learner; and react in immediate and helpful ways. On a more general level, we can support learning by using automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans (e.g., estimating competency levels across a network of skills, addressing what the person knows and can do, and to what degree). These *competency-level diagnoses* then provide the basis for improved instruction, self-reflection, and so on.

One critical problem we face is how to make sense of what can potentially become a deluge of information. What is wheat and what is chaff? My currently preferred solution involves using evidence-centered design (ECD), which supports both levels of diagnosis, and thus can be used for formative and summative purposes, and more importantly to enhance student learning (Mislevy, Steinberg, & Almond, 2003). It clarifies the "wheat" in performance data.

Stealth Assessment, Generally Speaking

Stealth assessment is seamlessly woven directly into the fabric of the instructional environment to support learning of important content and key competencies. This represents a quiet, yet powerful process by which learner performance data are continuously gathered during the course of playing/learning and inferences are made about the level of relevant competencies (see Shute, Ventura, Bauer, & Zapata-Rivera, 2009). Inferences on competency states are stored in a dynamic model of the learner. Stealth assessment is intended to support learning and maintain *flow*, defined as a state of optimal experience, where a person is so engaged in the activity at hand that self-consciousness disappears, sense of time is lost, and the person engages in complex, goal-directed activity not for external rewards, but simply

for the exhilaration of doing (Csikszentmihalyi, 1990). Stealth assessment is also intended to remove (or seriously reduce) test anxiety, while not sacrificing validity and reliability (Shute, Hansen, & Almond, 2008). The goal is to eventually blur the distinction between assessment and learning.

Key elements of the approach include: (1) evidence-centered assessment design, which systematically analyzes the assessment argument concerning claims about the learner and the evidence that supports those claims (Mislevy et al., 2003); and (2) formative assessment and feedback to support learning (Black & Wiliam, 1998a; 1998b; Shute, 2008). Additionally, stealth assessment provides the basis for instructional decisions, such as the delivery of tailored content to learners (e.g., Shute & Towle, 2003; Shute & Zapata-Rivera, 2008). Information is maintained within a learner model and may include cognitive as well as noncognitive information for a broader, richer profile.

During the last couple of years, my interest in this area has reignited, and my colleagues, students, and I have been working (playing) to firm up and flesh out the ideas relating to stealth assessment using computer-based games as our research vehicle of choice. We have been focusing on so-called 21st-century (or more simply "key") competencies (e.g., systems thinking, creative problem solving, identity management, teamwork, perspective taking, and time management). In conjunction with reviewing the literature on each of these competencies, we've been modeling, refining, and beginning to validate these emerging models via experts' reviews of the models. To test the viability of the models within immersive games, we have been exploring and analyzing various games/virtual worlds to use for test driving the models and ideas, and have built a couple of "worked examples" within some existing games. That is, we have recently modeled some key competencies within game environments, including (1) creative problem solving (within The Elder Scrolls IV: Oblivion, 2006, by Bethesda Softworks; Shute et al., 2009), and (2) systems thinking skills. Regarding the latter competency, we recently provided a worked example of an existing 3D immersive game called Quest Atlantis: Taiga Park (e.g., Barab, 2006; Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007). Both of these are described later in more detail.

Why Stealth Assessment Is Needed

What motivates this research to identify key competencies and use games as instructional learning vehicles? In a nutshell, the world is effectively shrinking and getting more complex. For instance, we're confronted with problems of enormous complexity and global ramifications (e.g., nuclear proliferation, global warming, antibiotic-resistant microbes, and

destruction of the rain forests). When faced with such complex problems, the ability to think creatively, critically, collaboratively, and systemically and then communicate effectively is essential. Learning and succeeding in a complex and dynamic world is not easily measured by multiple-choice responses on a simple knowledge test. Instead, solutions begin with rethinking assessment, identifying new skills and standards relevant for the 21st century, and then figuring out how we can best assess students' acquisition of the key competencies.

Currently, there is a large gap between what kids do for fun and what they're required to do in school. School covers material that is deemed "important," but kids are often unimpressed. These same kids, however, are highly motivated by what they do for fun (e.g., play games, participate in social networking sites). This mismatch between mandated school activities and what kids choose to do on their own is cause for concern regarding the motivational impact (or lack thereof) of school, but it needn't be the case. Imagine these two worlds united. Student engagement is strongly associated with academic achievement; thus, embedding school material within game-like environments has great potential to increase learning, especially for disengaged students.

The main assumptions underlying stealth assessment research are that: (1) learning by doing (required in game play) improves learning processes and outcomes; (2) different types of learning and learner attributes may be verified and measured during game play; (3) strengths and weaknesses of the learner may be capitalized on and bolstered, respectively, to improve learning; and (4) formative feedback can be used to further support student learning (Gee, 2003; Shute, 2007, 2008; Shute, Hansen, & Almond, 2008; Squire, 2006).

I now briefly define games, learning, and assessment in the context of this chapter. These definitions are followed by a section describing the evidence-based foundation on which stealth assessment rests.

BRIEF DEFINITION OF TERMS

Computer-Based Games

In their seminal book on the topic, *Rules of Play*, Salen and Zimmerman (2004) define a game as "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome" (p. 80). In addition to conflict, rules, and outcomes, Prensky (2001) adds goals, feedback, interaction, and representation (or story) into the mix of essential game elements. The combined list of essential game elements as used in this chapter includes: (1) conflict or challenge (i.e., a problem to be solved), (2)

rules of engagement, (3) particular goals or outcomes to achieve (which often include many sub-goals), (4) continuous feedback (mostly implicit, but may be explicitly cognitive and/or affective), (5) interaction within the environment, and (6) compelling story and representations. This inventory of important game elements is actually quite similar to those underlying good instructional design, but excludes design-free activities (e.g., make-believe games), where there are likely to be rules but unlikely to be quantifiable outcomes, such as points or rank accrued. Also note that this definition is parallel to the idea of assessment, with the purpose of describing knowledge, skills, and other attributes in a quantifiable manner.

Narrowing the definition a bit further, this chapter focuses on *interactive, digital games that support learning and/or skill acquisition.* This narrower definition is still pretty broad, and includes serious games as well as casual, educational, action, adventure, strategy, role-playing, puzzle, simulation, and massively multiplayer online games.

One reason why games are so engaging is because kids (of all ages) like to be in control of what's on the screen, and games offer this control on a continuing basis. In addition, games can give kids a powerful sense of mastery. Success is addictive, and computer-based games provide constant doses of small successes as players defeat more enemies, earn higher scores, and graduate to more challenging levels. In addition to fostering feelings of control and mastery, other reasons that games are so engaging are because players are motivated by social interaction, competition, knowledge, and escapism (Hirumi, Appelman, Rieber, & Van Eck, in press; Novak, 2005). Similarly, Prensky (2001) cites a number of ways that games capture and sustain players' interest including sensation, fantasy, narrative, fellowship, discovery, and expression. Once engaged, learning takes place naturally within the storyline of a well-designed game. The key, then, is seamlessly aligning "story" and "lesson"—a non-trivial endeavor (see Rieber, 1996).

Learning in Games

In general, learning is at its best when it is active, goal-oriented, contextualized, and interesting (e.g., Bransford, Brown, & Cocking, 2000; Bruner, 1961; Quinn, 2005; Vygotsky, 1978). Instructional environments should thus be interactive, provide ongoing feedback, grab and sustain attention, and have appropriate and adaptive levels of challenge—in other words, have the features of good games. Gee (2003) has persuasively argued that the secret of an immersive game as an instructional system is not its 3D graphics and other bells and whistles, but its underlying architecture. 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., Falmagne,

Cosyn, Doignon, & Thiery, 2003; Vygotsky, 1987) have long argued that the best instruction hovers at the boundary of a student's competence.

Recent reports (e.g., Thai, Lowenstein, Ching, & Rejeski, 2009) have further contended that well-designed games can act as *transformative digital learning tools* to support the development of skills across a range of critical educational areas. The simple logic mentioned earlier is that compelling storylines represent an important feature of well-designed games that tend to induce flow (Csikszentmihalyi, 1990), which in turn is conducive to learning. One major problem (as will be discussed in a later section) is that immersive games lack an assessment infrastructure to maximize learning potential. Furthermore, typical assessments are likely to disrupt flow in good games. Thus, there is a need for embedded (i.e., stealth) assessments that would be less obtrusive and hence less disruptive to flow.

Assessment in Games

In games, as players interact with the environment, the values of different game-specific variables change. For instance, getting injured in a battle reduces health, finding a treasure or other object increases your inventory of goods, and so on. In addition, solving major problems in games permits players to gain rank. One could argue that these are all "assessments" in games—of health, personal goods, and rank. But now consider including additional variables in games. Suddenly, in addition to checking health status, players could monitor their systems-thinking skills, creativity, and teamwork skills, and if values of those variables got too low, the player would likely take action to help boost them.

Playing well-designed games certainly has the potential to enhance learning, and more researchers every year are claiming that a lot of important learning and development is going on within such games (e.g., Dede, this volume; Green & Bavelier, 2003; Tobias & Fletcher, 2007). But what exactly is being learned? Are students/players learning what's intended via the game design? Are these skills educationally valuable (especially with an eye toward future workforce needs)? And how can we substantiate these claims? These questions are addressed in the following section on how to develop good stealth assessment.

The main challenge for educators who want to employ or design games to support learning is making valid inferences about what the student knows, believes, and can do without disrupting the flow of the game (and hence student engagement and learning). One solution entails the use of an assessment design approach called evidence-centered design (Mislevy, Steinberg, & Almond, 2003), which enables the estimation of students' competency levels and further provides evidence supporting claims about

competencies. Consequently, ECD has built-in diagnostic capabilities that allow any stakeholder (i.e., the teacher, student, parent, and others) to examine the evidence and view the current estimated competency levels. This in turn can inform instructional support.

The contribution of ECD to the story of measuring what students are getting from their interactions with games relates to its ability to equally and accurately assess lower- as well as higher-order thinking skills as distinguished in Anderson and Krathwohl's (2001) categorization (i.e., lower-level skills include knowledge, comprehension and application, while higher-level skills include analysis, synthesis, evaluation, and creation). Historically, higher-level skills, requiring students to think critically and creatively, are very difficult to assess, yet those skills are quite suitable for evidence-based stealth assessment.

Before describing the two stealth assessment examples, I now present the foundation underlying stealth assessment.

HOW TO DESIGN AND DEVELOP GOOD STEALTH ASSESSMENT

There are several problems that must be overcome to incorporate assessment in games. Bauer, Williamson, Mislevy, and Behrens (2003) address many of these same issues with respect to incorporating assessment within interactive simulations. In playing games, learners/players naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess (e.g., collaboration, critical thinking, problem solving). Evidence needed to assess the skills is thus provided by the players' interactions with the game itself (i.e., the processes of play), which may be contrasted with the product(s) of an activity, which is the norm within educational and training environments.

Making use of this stream of evidence to assess knowledge, skills, and understanding (as well as beliefs, feelings, and other learner states and traits) presents problems for traditional measurement models used in assessment. First, in traditional tests the answer to each question is seen as an independent data point. In contrast, the individual actions within a sequence of interactions in a simulation or game are often highly dependent on one another (e.g., Brown, Burton, & DeKleer, 1982). For example, what one does in a combat game at one point in time affects subsequent actions later on. Second, in traditional tests, questions are often designed to get at one particular piece of knowledge or skill. Answering the question correctly is evidence that one knows a certain fact: one question = one fact. But by analyzing responses to *all* of the questions or a sequence of actions (where each response or action provides incremental evidence about the current

mastery of a specific fact, concept, or skill), instructional environments may infer what learners are likely to know and not know overall.

Because we typically want to assess a whole cluster of skills and abilities from evidence coming from learners' interactions within a game or simulation, methods for analyzing the sequence of behaviors to infer these abilities are not as obvious. ECD is a method that can address these problems and enable the development of robust and valid simulation- or game-based learning systems. Bayesian networks comprise a powerful tool to accomplish these goals. ECD and Bayes networks are described in turn.

Evidence-Centered Design

The fundamental ideas underlying ECD came from Messick (1994) and then formalized by Mislevy and colleagues. This process begins by identifying what should be assessed in terms of knowledge, skills, or other attributes. These variables cannot be observed directly, so behaviors and performances that demonstrate these variables should be identified instead. The next step is determining the types of tasks or situations that would draw out such behaviors or performances. An overview of the ECD approach is described below (for more on the topic, see Mislevy & Haertel, 2006; Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, & Almond, 2003).

A game that includes stealth assessment must elicit behavior that bears evidence about key skills and knowledge, and it must additionally provide principled interpretations of that evidence in terms that suit the purpose of the assessment (Mislevy, Steinberg, & Almond, 2003). Working out these variables and models and their interrelationships is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design:

- *Competency Model: What collection of knowledge and skills should be assessed?* A given assessment is meant to support inferences for some purpose, such as grading, certification, diagnosis, guidance for further instruction, and so on. Variables in the competency model (*CM*) are usually called *nodes* and describe the set of knowledge and skills on which inferences are to be based. The term *student model* is used to denote a student-instantiated version of the CM—like a profile or report card, only at a more refined grain size. Values in the student model express the assessor's current belief about a learner's level on each variable in the CM.
- *Evidence Model: What behaviors or performances should reveal those constructs?* An evidence model expresses how the learner's interactions with and responses to a given problem constitute evidence about

competency model variables. The evidence model (EM) attempts to answer two questions: (1) What behaviors or performances reveal targeted competencies? and (2) What is the functional (or statistical) connection between those behaviors and the CM variable(s)? Basically, an evidence model lays out the argument about why and how the observations in a given task situation (i.e., learner performance data) constitute evidence about CM variables.

- *Task Model: What tasks should elicit those behaviors that comprise the evidence?* Task-model variables, used in typical assessment design, describe features of situations that will be used to elicit performance. A task model (TM) provides a framework for characterizing and constructing situations with which a student will interact to provide evidence about targeted aspects of knowledge related to competencies. Task specifications establish what the learner will be asked to do, what kinds of responses are permitted, what types of formats are available, and so on. Tasks are the most obvious part of an assessment, and their main purpose is to elicit evidence (which is observable) about competencies (which are unobservable). For stealth assessment in games, I use the term "action model" instead of task model. This reflects the fact that we are dynamically modeling students' *action sequences*. These action sequences form the basis for drawing evidence and inferences and may be compared to simpler task responses as with typical assessments. The action model in a gaming situation defines the sequence of actions and each action's indicators of success. Actions represent the things that students do to complete the mission or solve a problem.

In games with stealth assessment, the student model accumulates and represents belief about the targeted aspects of skill, expressed as probability distributions for competency-model variables (Almond & Mislevy, 1999). Evidence models identify what the student says or does that can provide evidence about those skills (Steinberg & Gitomer, 1996) and express in a psychometric model how the evidence depends on the competency-model variables (Mislevy, 1994). Task/action models express situations that can evoke required evidence. One effective tool that I've been employing in various competency and evidence modeling efforts is Bayesian networks.

Bayesian Networks

Bayesian networks (Pearl, 1988) may be used within student models to handle uncertainty by using probabilistic inference to update and improve belief values (e.g., regarding learner competencies). The inductive and de-

ductive reasoning capabilities of Bayesian nets support "what-if" scenarios by activating and observing evidence that describes a particular case or situation, and then propagating that information through the network using the internal probability distributions that govern the behavior of the Bayesian net. Resulting probabilities inform decision making, as needed in, for instance, the selection of the best chunk of content or instructional support to subsequently deliver based on the learner's current state. (Examples of Bayes net implementations for student models may be seen in Conati, Gertner, & VanLehn, 2002; Shute, Graf, & Hansen, 2005; VanLehn et al., 2005.)

EXAMPLES OF STEALTH ASSESSMENT SYSTEMS

Thinking in Taiga Park—Example 1

In the first example, I focus on systems thinking as a key competency worthy of support for success in the 21st century. The game I selected for the worked example is called Taiga Park, an immersive, 3D role-playing game helping middle-school kids to learn important knowledge and skills related to ecology and scientific inquiry. Taiga Park features a beautiful virtual park with a river running through it (Barab, Zuiker, et al., 2007; Zuiker, 2007). The park is populated by several groups of people who use or depend on the river in some capacity. Although the groups are quite different, their lives (and livelihoods) are entwined, demonstrating several levels of "systems" within the world (e.g., the ecological system comprising the river and the socio-economic system comprising the groups of stakeholders in the park). In addition to the park ranger (Ranger Bartle), the three stakeholders include: (1) the Mulu (indigenous) farmers, (2) Build-Rite Timber Company, and (3) the K-Fly Fishing Tour Company. There are also park visitors, lab technicians, and others with their own sets of interests and areas of expertise.

The Taiga storyline is about how the fish population in the Taiga River is dying. Students participate in this world by helping Ranger Bartle figure out how he can solve this problem of the declining fish population and thus save the park. Students begin the series of five missions by reading an introductory letter from Ranger Bartle. In the letter, Ranger Bartle pleads for help and states his need for an expert field investigator (i.e., you, the player/student) who can help him solve the declining fish population problem.

The rationale for using systems thinking as the focal competency is that problems facing today's citizens (e.g., healthcare reform, the need for new energy sources independent of fossil fuels, a plastic island the size of Texas in the Pacific, and persistent racial and religious intolerance) are complex, dynamic, and cannot be solved unilaterally. Furthermore, many of these

problems are ill-structured in that there is not just one correct solution. Instead, we need to think in terms of the underlying system and its subsystems to solve these kinds of problems (Richmond, 1993). The ability to act effectively in such complex situations requires competence in what's called systems thinking (ST) skill (Arndt, 2006).

To accomplish stealth assessment of systems thinking in a real game environment, my students and I began by developing a competency model relevant to systems thinking skill (see Figure 20.1).

For the worked example, we focused on just one branch of the CM: "model the system" (which includes all the shaded nodes connected to the right of it). Our systems thinking CM was created after an extensive literature review on the topic. Nodes in the CM were statistically linked to each other in terms of conditional probabilities and comprise different levels in the network. For instance, the "parent" node represents an estimate of the learner's general systems thinking skill, given all of the evidence collected at that point. This is a latent, unobservable construct, as is the "model the system" node. Low-level nodes (i.e., those without progeny) are explicit-

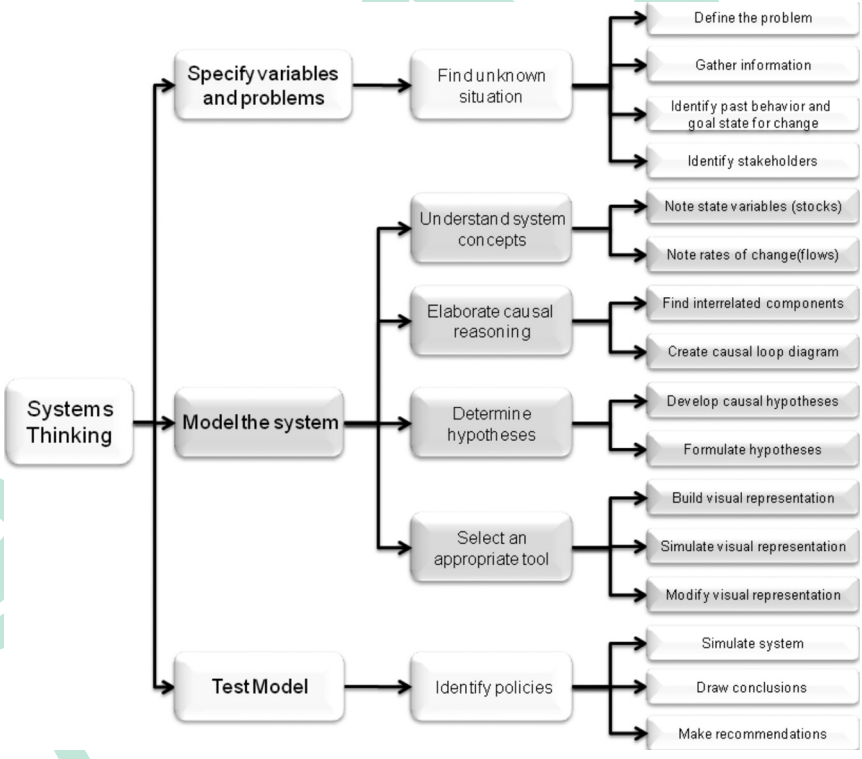

Figure 20.1 Competency model for systems thinking skill.

ly linked to indicators (observable behaviors/actions) obtained from the game via our evidence model. Such indicators provide information that "feeds" the Bayes net. For instance, one low-level node in Figure 20.1 is "gather information." In the Taiga Park world, several quests require the player to demonstrate that skill, with indicators relating to the accuracy and efficiency of doing so. One example task requires the student to collect water samples at various spots along the river. Another requires the student to take photographs of the river at different locations and times. Successful completion of those specific indicators provides data for the "gather information" node. Once the information is inserted into the Bayes net, it is propagated throughout the network to all of the nodes, whose estimates are subsequently updated.

The quests that players undertake in Taiga Park all relate to solving the overarching problem of a rapidly declining fish population on which three major groups of Taiga Park stakeholders rely (i.e., native farmers, loggers, and a fishing tournament company). Quests take place within five different "Missions," all of which are designed to make learners think carefully about complex ecological systems—their interconnections and dynamic relations among elements. Thus, the fit between our selected competency and that goal of the game was ideal.

As part of the worked example, we created a Bayesian network for a subset of the CM—model the system. We then modeled a hypothetical learner (Clara) in terms of her systems thinking skill at two points in time: an initial quest (Time 1) and a final one (Time 2). The example showed quantitative and qualitative changes to her systems thinking over time. For instance, we compared Clara's causal loop diagrams—depicting current understanding of factors causing the fish to die—created at Time 1 and Time 2 to an expert's diagram (note: "Create a causal loop diagram" is on the far right of Figure 20.1, 8th node down). These comparisons are made possible by automatically standardizing her diagram, and then overlaying the standardized map onto an expert map. The tool that we used for the standardization and comparison is an Excel-based software application called jMap (Jeong, 2008; Shute, Jeong, & Zapata-Rivera, in press) was designed to accomplish the following goals: (1) elicit, record, and automatically code mental models; (2) visually and quantitatively assess changes in mental models over time; and (3) determine the degree to which the changes converge towards an expert's (for more information about the program as well as relevant paper and links, see: http://garnet.fsu.edu/~ajeong).

Information that is obtained from comparing Clara's causal diagram to an expert map (a) provides input to the Bayes net relating to that node, and (b) clearly demonstrates any misconceptions which can be used as the basis for formative feedback presented to the learner by the teacher or automatically by the environment. For example, given particular errors

of omission apparent in Clara's early map,[1] the system would provide the following feedback "Nice job, Clara—but you forgot to include the fact that sediment increases water temperature which decreases the amount of dissolved oxygen in the water. That's the reason the fish are dying—they don't have enough oxygen." For graphical feedback, the Taiga lab technician (or another knowledgeable character in the park) could give Clara the expert causal loop diagram, explicitly highlighting her omitted variables in the picture. That way, she could see for herself what she'd left out. For more details, see Shute, Masduki, and colleagues (in press).

Creative Problem Solving in Oblivion—Example 2

Oblivion (The Elder Scrolls IV: Oblivion, 2006, by Bethesda Softworks) is the name of a commercial game with no pretense of being "educational." This is a first-person, 3D, role-playing game set in a medieval world. Upon entering the world, you choose to be one of many characters (e.g., knight, mage, elf), each of whom has (or can obtain) various weapons, spells, and tools. Your primary goal is to gain rank and complete quests, as with most of the games of this type. Quests may include locating a person to obtain information, figuring out a clue for future quests, and so on. There are multiple mini-quests along the way, and a major quest that results in winning the game. Players have the freedom to complete quests in any order they choose, and this can entail hundreds of hours of game play to complete the game.

The focus of this example is modeling and assessing creative problem solving (CPS). The simplified CM is shown in Figure 20.2, which includes some additional and educationally relevant competencies that might be assessed during game play in Oblivion. The shaded variables are used in this example. The CM, with its "cognitive" and "noncognitive" variables, should be viewed as illustrative only.

The evidence model defines the connections between specific observables and their underlying competencies. Observables are actions that are "scored" in relation to novelty and efficiency (indicators). The evidence model includes (1) scoring rules for extracting observables from students' game play indicators found in log files, (2) the observables (i.e., scored data), and (3) measurement rules for accumulating evidence from the observables, which are then used to update the student model variables. For simplicity, this illustration includes just two observables, each informing novelty or efficiency. Both of these, in turn, inform the CPS variable through intermediate variables (i.e., problem solving and creativity). The degree to which variables differentially inform their parent nodes is represented in a Bayes net.

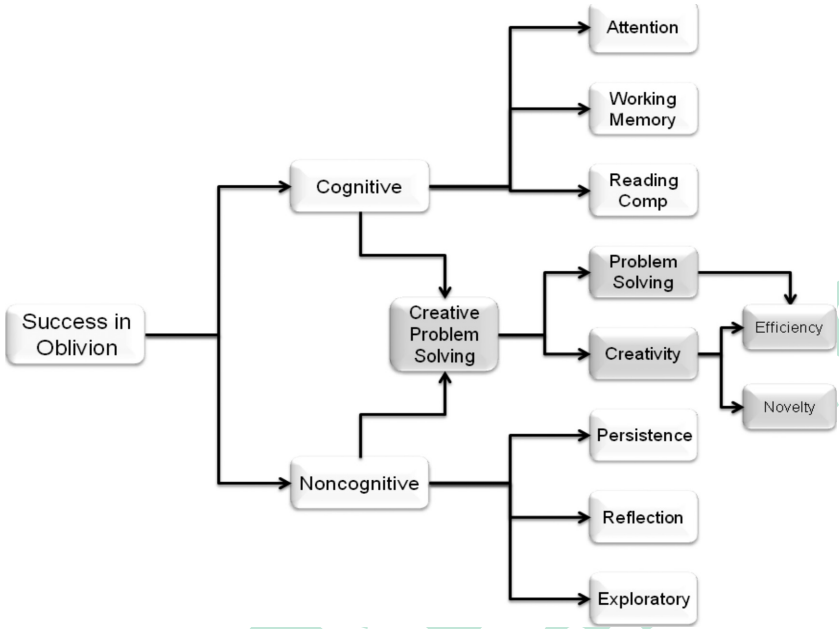

Figure 20.2 Competency model for creative problem solving.

Now, suppose you are faced with the problem of having to cross a river full of dangerous fish in Oblivion. There is a mage in a cave on the other side who has some information you need. Table 20.1 contains a list of actions to solve this problem, as well as the indicators that may be learned from real student data, or elicited from experts. For the system to learn indicator values from real data, estimates of *novelty,* for example, may be defined in terms of the frequency of use across all players. For instance, swimming across the river is a high-frequency, common solution, thus associated with a low "novelty weight." An estimate of *efficiency* may be defined in terms of the probability of successfully solving a problem given a set of

TABLE 20.1 Example of Action Model with Indicators for Novelty and Efficiency

Action	Novelty	Efficiency
Swim across the river	$n = 0.12$	$e = 0.22$
Levitate over the river	$n = 0.33$	$e = 0.70$
Freeze the river with a spell and slide across	$n = 0.76$	$e = 0.80$
Find a bridge over the river	$n = 0.66$	$e = 0.24$
Dig a tunnel under the river	$n = 0.78$	$e = 0.20$

actions—based on time and resources expended. Swimming across the river would thus have a low efficiency value because of the extra time needed to evade the piranha-like fish that live there. On the other hand, digging a tunnel under the river to get to the other side is judged as highly novel, but less efficient than, say, freezing the water and simply sliding across—the latter being highly novel and highly efficient. The indicator values shown in Table 20.1 were obtained from two *Oblivion* experts, and they range from 0 to 1. Higher numbers relate to greater levels of both novelty and efficiency. The two experts were very similar in their estimates of indicator values, and the values in the table represent an average of their estimates.

Actions can be captured in real time as the player interacts with the game, and associated indicators can be used to provide evidence for the appropriate competencies. This is accomplished via the evidence model using Bayesian network software. Figure 20.3 shows a Bayes net after a player elected to cross the river by digging a tunnel under it.

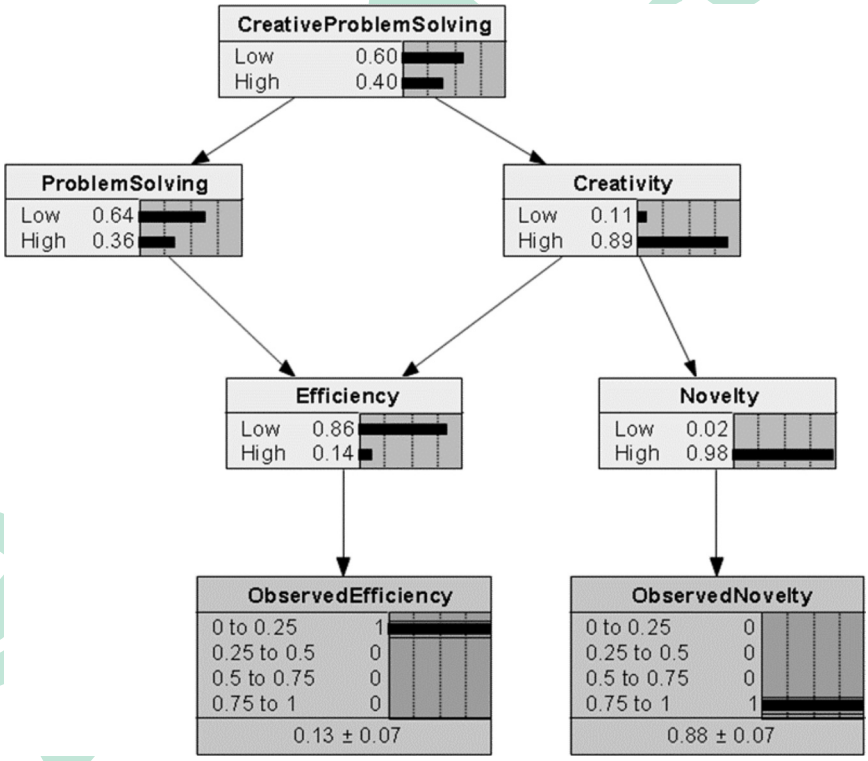

Figure 20.3 Bayes model showing marginal probabilities after observing a low efficiency and high novelty action of crossing the river by digging a tunnel under it.

We can see that even though the player evidenced very high novelty in her solution, the parent node of CPS is still inferring that she is more "low" than "high" on this attribute—illustrating that efficiency is a more valued competency than novelty, based on the way the CM was set up, and that she has many more chances to improve this skill during game play.

DISCUSSION

The challenge for educators who want to employ games to support learning is making valid inferences about what the student knows and can do without disrupting the flow of the game (and hence student engagement and learning). My solution entails the use of ECD, which enables the estimation of students' competency levels and further provides the evidence supporting claims about competencies. Consequently, ECD has built-in diagnostic capabilities that permit a stakeholder (i.e., the teacher, student, parent, and others) to examine the evidence and view the current estimated competency levels. This in turn can inform instructional support.

So what are some of the downsides of this approach? Implementing ECD within gaming environments poses its own set of challenges. For instance, Rupp, Gushta, Mislevy, and Shaffer (in press) have highlighted several issues that must be addressed when developing games that employ ECD for assessment design. The competency model, for example, must be developed at an appropriate level of granularity to be implemented in the assessment. Too large a grain size means less specific evidence is available to determine student competency, while too fine a grain size means a high level of complexity and increased resources to be devoted to the assessment. In addition, developing the evidence model can be rather difficult in a gaming environment when students collaborate on completing quests. For example, how would you trace the actions of each student and what he or she is thinking when the outcome is a combined effort? Another challenge comes from scoring qualitative products such as essays, student reflections, and online discussions where there remains a high level of subjectivity even when teachers are provided with comprehensive rubrics. Thus a detailed and robust coding scheme is needed that takes into account the context of the tasks and semantic nuances in the students' submissions. Finally, for the task or action model, issues remain in terms of how the assigned tasks should be structured (or not). While examining particular sequences of actions (e.g., as in Taiga Park) can facilitate more reliable data collection, it might limit the students' ability to explore the environment or go down alternative paths that make games more interesting and promote self-learning. Therefore, when game designers build assessments into the game, they

need to find the ideal balance between student exploration and structured data collection.

How do teachers fit into this effort? In games designed for educational purposes (like Taiga Park and unlike Oblivion), the system can allow teachers to view their students' progress during their missions via a web-based toolkit (note: in Taiga Park, this exists as the Teachers Toolkit panel). This lets teachers receive and grade all of the student submissions (which, across the various missions, may start to feel like a deluge). In our worked example, instead of spending countless hours grading essays and diagrams, teachers instead could review students' competency models, and use that information as the basis for altering instruction or providing formative feedback (see Shute, 2008). For example, if the competency models during a mission showed evidence of a widespread misconception, the teacher could turn that into a teachable moment, or may choose to assign struggling students to team up with more advanced students in their quests.

Information about students' competencies may also be used by the system to select new gaming experiences (e.g., more challenging quests could be made available for students who exhibit high CPS abilities). In addition, and as suggested earlier, up-to-date estimates of students' competencies, based on assessment information handled by the Bayes nets, can be integrated into the game and explicitly displayed as progress indicators. Players could then see how their competencies are changing based on their performance in the game. Most games already include status bars, representing the player's current levels of game-related variables. Imagine adding high-level competency bars that represent attributes like creative problem solving and systems thinking skill. More detailed information could be accessed by clicking the bar to see current states of lower-level variables. And like health status, if any competency bar gets too low, the student needs to act to somehow increase the value. Once students begin interacting with the bars, metacognitive processes may be enhanced by allowing the player to see game- or learning-related aspects of their state. Viewing their current competency levels and the underlying evidence gives students greater awareness of personal attributes. In the literature, these are called "open student models," and they have been shown to support knowledge awareness, reflection, and learning (Bull & Pain, 1995; Hartley & Mitrovic, 2002; Kay, 1998, Zapata-Rivera & Greer, 2004; Zapata-Rivera, Vanwinkle, Shute, Underwood, & Bauer, 2007).

Future research plans include formally implementing some of our stealth assessment examples directly into games (e.g., Taiga Park) to test the efficacy of the approach in relation to supporting students as well as teachers. We also have some other worked examples ready to go in relation to (1) assessing and supporting creative problem solving in Media Village (another Quest Atlantis world) and (2) assessing and supporting perspective taking

in Mesa Verde (part of the Quest Atlantis world). Some upcoming research challenges include figuring out how to ensure portability/transfer of competency models across games/environments, and how to model context in the student-learning picture.

In conclusion, the ideas in this chapter relate to using ECD, stealth assessment, and automated data collection and analysis tools to not only collect valid evidence of students' competency states and support student learning, but also to reduce teachers' workload in relation to managing the students' work (or actually "play") products. This would allow teachers to focus their energies on the business of fostering student learning. If educational games were easy to employ and provided integrated and automated assessment tools as described herein, then teachers would more likely want to utilize them to support student learning across a range of educationally valuable skills. The ideas and tools within this chapter are intended to help teachers facilitate learning, in a fun and engaging manner, of educationally valuable skills not currently supported in school.

NOTE

1. In this case, Clara blamed the loggers and created a diagram showing that: as park income went down, logging operations increased, which meant more trees were cut down, thus more sediment got into the water, causing more fish to die, and reducing income further, causing more logging to occur, in a looping cycle.

REFERENCES

Almond, R. G. & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 223–237.

Anderson, L. W. & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Addison-Wesley Longman.

Arndt, H. (2006). Enhancing system thinking in education using system dynamics. *Simulation, 82*(11), 795–806.

Barab, S. A. (2006). From Plato's Republic to Quest Atlantis: The role of the philosopher-king. *Technology, Humanities, Education, and Narrative, 2,* 22–53.

Barab, S. A., Sadler, T. D., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: Supporting consequential play. *Journal of Science Education and Technology, 16*(1), 59–82.

Barab, S. A., Zuiker, S., Warren, S., Hickey, D. Ingram-Goble, A, Kwon, E-J, Kouper, I, & Gerring, S. C. (2007). Situationally embodied curriculum: Relating formalisms and contexts. *Science Education, 91*(5), 750–782.

Bauer, M., Williamson, D., Mislevy, R. & Behrens, J. (2003). Using evidence-centered design to develop advanced simulation-based assessment and training. In G. Richards (Ed.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003* (pp. 1495–1502). Chesapeake, VA: AACE.

Bethesda Softworks (2006). *Elder Schools VI: Oblivion*. Retrieved from http://www.bethsoft.com/eng/games/games_oblivion.html

Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7–74.

Black, P. & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: School of Education, King's College.

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

Brown, J. S., Burton, R. R., & DeKleer, J. (1982). Pedagogical, natural language, and knowledge engineering in SOPHIE I, II, and III. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 227–282).

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21–32.

Bull, S. & Pain, H. (1995). "Did I say what I think I said, and do you agree with me?": Inspecting and questioning the student model. In *Proceedings of the Artificial Intelligence in Education* (pp. 501–508). Charlottesville, VA: AACE.

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling & User-Adapted Interaction, 12*(4), 371-417.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optical experience*. New York: Harper Perrennial.

Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiery, N. (2003). The assessment of knowledge, in theory and in practice. In R. Missaoui & J. Schmid (Eds.), ICFCA, Vol. 3874 of Lecture Notes in Computer Science, pp. 61–79. Springer.

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

Green, C. & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*, 534–537.

Hartley, D. & Mitrovic, A. (2002) Supporting Learning by opening the Student Model. In *Proceedings of ITS 2002*, pp. 453–462.

Hsumi, A., Appelman, R., Rieber, L., & Van Eck, R. (in press). Four perspectives on preparing instructional designers to optimize game-based learning. *Tech Trends*, AECT.

Jeong, J. C. (2008). *Discussion Analysis Tool* (DAT). Retrieved from http://garnet.fsu.edu/~ajeong/DAT

Kay, J. (1998). *A scrutable user modelling shell for user-adapted interaction*. Ph.D. Thesis, Basser Department of Computer Science, University of Sydney, Sydney, Australia.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher, 32*(2), 13–23.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483

Mislevy, R. J. & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). CA: Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED483399)

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.

Novak, J. (2005). *Game development essentials: An introduction.* New York: Thomson Delmar Learning.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Kaufmann.

Prensky, M. (2001). *Digital game-based learning.* New York: McGraw-Hill.

Quinn, C. (2005). *Engaging learning: Designing e-learning simulation games.* San Francisco: Pfeiffer.

Rieber, L. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Education and Technology Research & Development, 44,* 42–58.

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review, 9*(2), 113–133.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (in press). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment.*

Salen, K. & Zimmerman, E. (2004). *Rules of play: Game design fundamentals.* Cambridge, MA: MIT Press.

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139–187). Mahwah, NJ: Lawrence Erlbaum.

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

Shute, V. J. & Glaser, R. (1990). Large-scale evaluation of an intelligent tutoring system: Smithtown. *Interactive Learning Environments, 1,* 51–76.

Shute, V. J. & Glaser, R. (1991). An intelligent tutoring system for exploring principles of economics. In R. E. Snow & D. Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (pp. 333–366). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shute, V. J. & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist, 38*(2), *105–114.*

Shute, V. J. & Zapata-Rivera, D. (2008). Adaptive technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology* (3rd Edition) (pp. 277–294). New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J., Glaser, R. & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and Individual Differences* (pp. 279–326). New York: W.H. Freeman.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. PytlikZillig, R. Bruning, and M. Bodvarsson (Eds.). *Technology-based education: Bringing researchers and practitioners together* (pp. 169–202). Greenwich, CT: Information Age Publishing.

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., & Jeong, A., C., & Zapata-Rivera, D. (in press). Using flexible belief networks to assess mental models. In B. B. Lockee, L. Yamagata-Lynch, and J. M. Spector (Eds.), *Instructional Design for Complex Learning.* New York: Springer.

Shute, V. J., Masduki, I., Donmez, O., Kim, Y. J., Dennen, V. P., Jeong, A. C., & Wang, C-Y. (in press). Assessing key competencies within game environments. To appear in D. Ifenthaler, P. Pirnay-Dummer, N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge*, New York: Springer-Verlag.

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.

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

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24,* 223–258.

Thai, A., Lowenstein, D., Ching, D., & Rejeski, D. (2009). *Game changer: Investing in digital play to advance children's learning and health.* New York: The Joan Ganz Cooney Center at Sesame Workshop.

Tobias, S. & Fletcher, J. D. (2007). What research has to say about designing computer games for learning. *Educational Technology, 47*(5), 20–29.

Van Lehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L. Treacy, D. Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned, *International Journal of Artificial Intelligence and Education, 15*(3), 1–47.

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.

Zapata-Rivera, D. & Greer, J. E. (2004). Interacting with inspectable Bayesian models. *International Journal of Artificial Intelligence in Education, 14*, 127–163.

Zapata-Rivera, D., Vanwinkle, W., Shute, V. J., Underwood, J. S., & Bauer, M. (2007). English ABLE. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education–Building technology rich learning contexts that work* (pp. 323–330). Amsterdam: IOS Press.

Zuiker, S. (2007). *Transforming practice: Designing for liminal transitions along trajectories of participation.* Unpublished doctoral dissertation. Indiana University, Bloomington, Indiana.

Author Queries:

Is this a quote from someone: “dances around the outer limits of the player’s abilities,” (from manuscript p. 7)? If so, please give the source of the quote. If not, the quotation marks are a little confusing.

In your references, please give publisher information for Brown, Burton, & DeKleer (1982) and for Hartley & Mitrovic (2002)

Please update (if applicable) the publication status of Hirumi et al.; Rupp et al.; Shute, Jeong, et al.; and Shute, Masduki, et al. (in press).

Please give the city along with the state for the location for Misleavy, Almond, & Lukas (2004).