Chapter 15
Modeling, Assessing, and Supporting Key Competencies Within Game Environments

Valerie J. Shute, Iskandaria Masduki, Oktay Donmez, Vanessa P. Dennen, Yoon-Jeon Kim, Allan C. Jeong, and Chen-Yen Wang

15.1 Introduction

> Human beings, viewed as behaving systems, are quite simple. The apparent complexity of our behavior is largely a reflection of the complexity of the environment in which we find ourselves. (Herbert A. Simon, 1996, p. 53)

A critical challenge for any successful instructional-learning system involves accurately identifying characteristics of a particular learner or group of learners – such as the type and level of specific knowledge, skills, and other attributes. This information can then be used to improve subsequent learning (Conati, 2002; Park & Lee; 2003; Shute, Lajoie, & Gluck, 2000; Snow, 1994). But *what* are the most valuable competencies needed to succeed in the twenty-first century, and *how* can we assess them accurately and support their development? These questions comprise the crux of our research, with a focus on the "how" part of the story in this chapter.

To put our research issues in context, the demands associated with living in a highly technological and globally competitive world require today's students to develop a very different set of skills than their parents (and grandparents) needed. That is, when society changes, the skills that citizens need to negotiate the complexities of life also change. In the past, a person who had acquired basic reading, writing, and calculating skills was considered to be sufficiently literate. Now, people are expected to read critically, write persuasively, think and reason logically, and solve increasingly complex problems in math, science, and everyday life. The general goal of education is to prepare young people to live independent and productive lives. Unfortunately, our current educational system is not keeping pace with these changes and demands of today's more complex environment.

V.J. Shute (✉)
Florida State University, Tallahassee, FL, USA
e-mail: shute@mail.coe.fsu.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_15,

15.1.1 Purpose

This chapter will describe our ideas and tools for modeling, assessing, and supporting key competencies (e.g., systems thinking, creativity, and collaboration) via formative assessment embedded within immersive games. Through an extensive literature review described elsewhere (Shute, Dennen, Kim, Donmez, & Wang, 2008), we have identified and have begun modeling a set of educationally valuable attributes, or *competencies*, that are currently being ignored in our schools (locally and globally), but we believe should not be – especially with an eye toward the near future. Our modeling efforts extend an existing evidence-centered design (ECD) approach formulated by Mislevy, Steinberg, and Almond (2003) and employ Bayesian networks (Pearl, 1988). That is, inferences – both diagnostic and predictive – are handled by Bayes nets and used directly in the student models to handle uncertainty via probabilistic inference to update and improve belief values on learner competencies. To make these ideas more concrete, we present an analysis (or worked example) of an existing 3D immersive game called Quest Atlantis: *Taiga Park* (e.g., Barab, 2006; Barab, Zuiker et al., 2007; Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007), and demonstrate how evidence is gathered and interpreted in relation to one of our targeted competencies: systems thinking skill.

The longer term goal of our research, outside the scope of this chapter, is to fully develop, refine, pilot test, and ultimately validate our evidence-based approach using stealth assessment embedded within immersive learning environments (e.g., games, simulations, scenarios) that can elicit data from learners, make inferences about competency levels at various grain sizes, and use that information as the basis for targeted and immediate support. The motivation for this research is the belief that certain attributes of people, such as insulating against opposing views, reducing complex issues to black-and-white terms, and failing to question entrenched ideas will likely *not* move us – citizens of the world – in the direction necessary to flourish in the twenty-first century. Our research goals are toward ensuring that current and future *worldizens* can learn to systematically and creatively think, communicate, question, collaborate, solve difficult problems, reflect on decisions and solutions to problems, and adapt to rapidly changing circumstances.

There are many obstacles that need to be overcome before education is truly effective for the future and for the masses (e.g., shortage of well-qualified teachers, inadequate financial resources for poor schools, delivery of content in ways that do not engage students, reliance on tests to get numbers instead of insight). One obstacle that is not usually included in the various lists – but should be – concerns a lack of clear vision about what exactly we are preparing our kids for. We can readily identify trends, such as the *shrinking world* phenomenon that occurs as we become progressively more interconnected. And we know that in the long run, it is less important to memorize information than to know how to locate and make sense of credible information. But do our schools alter their curricula to accommodate these emergent needs? No. Are we adequately preparing our students for the realities of their future? No. Students are still pushed to memorize and repeat facts, and consequently they are graduating high school ill-prepared to tackle real-world, complex problems. We cannot directly adjust the wind (the future), but we *can* adjust the sails (competencies). To do so effectively, we need to have a good sense of bearings – where we are, and where we are heading.

15.1.2 Where We Are

This section briefly overviews two major problems confronting us today: (a) disengaged students, and (b) an effectively shrinking world, commensurate with increased communication technologies (e.g., Barab, Zuiker et al., 2007; Gee, 2004a, 2004b; Shute, 2007). It provides the basic rationale for our moving toward authentic, engaging learning activities and related stealth assessment to support learning.

15.1.2.1 Disengaged Students

There is a huge gulf between what kids do for fun and what they are required to do in school. School covers material that we deem important, but kids, generally speaking, are unimpressed. These same kids, however, are highly motivated by what they do for fun (e.g., play interactive games). 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 need not be the case. Imagine these two worlds united. Student engagement is strongly associated with academic achievement; thus, combining school material with games has tremendous potential to increase learning, especially for lower performing, disengaged students. The logic underlying the research is as follows. Compelling storylines (narratives) represent an important feature of well-designed games. Well-designed games tend to induce *flow* (Csikszentmihalyi, 1990), a state in which a game player loses track of time and is absorbed in the experience of game play. Flow is conducive to engagement, and engagement is conducive to learning. The problem 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.

15.1.2.2 The Shrinking World

The second problem motivating our research is that the world is effectively shrinking. We are currently confronted with problems of enormous complexity and global ramifications (e.g., the massive meltdown on Wall Street, nuclear proliferation, global warming, a plastic island the size of Texas in the Pacific, antibiotic resistant microbes, destruction of the rain forests, and poverty). The people who will be making and managing policy decisions in the near future need to be able to understand, at the very least, how research works and how science works because solutions are going to be highly technical and highly complex. When confronted by problems, especially new issues for which solutions must be created out of whole cloth, the ability to think creatively, critically, collaboratively, 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 re-thinking assessment, identifying new skills and standards relevant for the twenty-first century, and then figuring out how we can best assess students' acquisition of the new competencies – which may in fact involve the teacher, the computer, the student, one's peers, and so on. Moreover, the envisioned new competencies should include not only cognitive variables (e.g., critical thinking and reasoning skills) but also noncognitive variables (e.g., teamwork, tolerance, and tenacity) as the basis for new assessments to support learning (Abedi & O'Neil, 2005; Farkas, 2003).

15.1.3 Where We Should Be Heading

The primary goal of this chapter is to figure out *how* to accomplish the design and development of valid and reliable assessments for critical competencies. As a preliminary step, we have begun to identify key competencies (see Fig. 15.1). This is not a comprehensive list; additional competencies will be identified and modeled as our research evolves. In this chapter we will model systems thinking skill to demonstrate how evidence-based assessments might be developed and embedded within games and simulation environments. Modeling, assessing, and supporting students in relation to our set of skills is intended to allow students to grow in a number of important new areas, function productively within multidisciplinary teams, identify and solve problems (with innovative solutions), and communicate effectively.

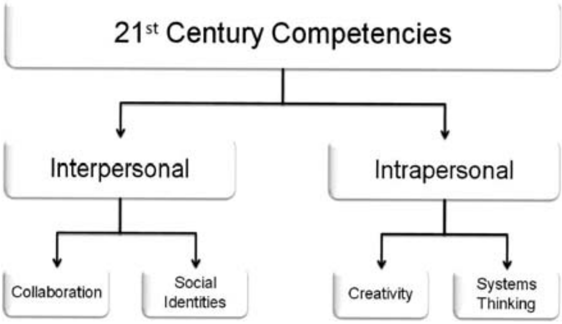


Fig. 15.1 Current set of key competencies for the twenty-first century

To accomplish our goal of developing really good assessments that can also support learning, we turn now to the "how" part of the story; namely, an overview of evidence-centered design (ECD) which supports the design of valid assessments. ECD entails developing competency models and associated assessments. We extend ECD by embedding these evidence-based assessments within interactive environments – comprising stealth assessment. Afterward, we present (a) a literature review and comprehensive model associated with the systems thinking competency and (b) a description of how these ideas would actually play out within an existing immersive game – Quest Atlantis: Taiga Park.

15.2 Assessment Methodology: Evidence-Centered Design

> The nature of the construct being assessed should guide the selection or construction of relevant tasks, as well as the rational development of construct-based scoring criteria and rubrics. (Sam Messick, 1994, p. 17)

The fundamental ideas underlying ECD came from Messick (1994; see quote above). 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 et al., 2003).

15.2.1 ECD Models

The primary purpose of an assessment is to collect information that will enable the assessor to make inferences about students' competency states – what they know, believe, and can do, and to what degree. Accurate inferences of competency states support instructional decisions that can promote learning. ECD defines a framework that consists of three theoretical models that work in concert. The ECD framework allows/requires an assessor to: (a) define the claims to be made about students' competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of tasks that will elicit that evidence. These three actions map directly onto the three main models of ECD shown in Fig. 15.2.

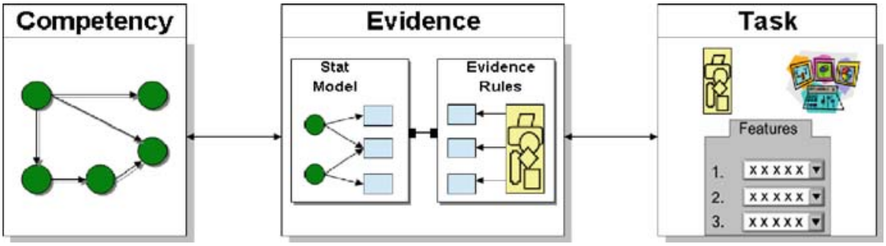


Fig. 15.2 Three main models of an evidence-centered assessment design

A good assessment has to elicit behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence in terms that suit the purpose of the assessment. Working out these variables, models, and their interrelationships is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design.

15.2.1.1 Competency Model

What collection of knowledge, skills, and other attributes should be assessed? This can also be phrased as: What do you want to say about the person at the end of the assessment? Variables in the competency model (*CM*) are usually called "nodes" and describe the set of person variables on which inferences are based. The term "student 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 student's level on each variable within the CM. For example, suppose the CM for a science class that valued the general competency of systems thinking contained a node for "Create a causal loop diagram." The value of that node – for a student who was really facile at understanding and drawing causal loop diagrams – may be "high" (if the competency levels were divided into low, medium, and high), based on evidence accumulated across multiple, relevant tasks.

15.2.1.2 Evidence Model

What behaviors or performances should reveal differential levels of the targeted competencies? An evidence model expresses how the student's interactions with, and responses to a given problem constitute evidence about competency model variables. The evidence model (EM) attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies; and (b) What's the connection between those behaviors and the CM variable(s)? Basically, an evidence model lays out the argument about why and how observations in a given task situation (i.e., student performance data) constitute evidence about CM variables. Using the same node as illustrated in the CM section above, the evidence model would clearly indicate the aspects of causal loop diagrams that must be present (or absent) to indicate varying degrees of understanding or mastery of that competency. The same logic/methods apply to noncognitive variables as well – stating clearly the rubrics for scoring aspects of creativity, teamwork, etc.

15.2.1.3 Task Model

What tasks should elicit those behaviors that comprise the evidence? 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 or skill related to competencies. These situations are described in terms of: (a) the presentation format (e.g., directions, stimuli), (b) the specific work or response products (e.g., answers, work samples), and (c) other variables used to describe key features of tasks (e.g., knowledge type, difficulty level). Thus, task specifications establish what the student will be asked to do, what kinds of responses are permitted, what types of formats are available, and other considerations, such as whether the student will be timed, allowed to use tools (e.g., calculators, dictionaries), and so forth. Multiple task models can be employed in a given assessment. Tasks are the most obvious part of an assessment, and their main purpose is to elicit evidence (which is observable) about competencies (which are unobservable).

15.2.1.4 Design and Diagnosis

As shown in Fig. 15.2, assessment design flows from left to right, although in practice it is more iterative. Diagnosis (or inference) flows in the opposite direction. That is, an assessment is administered, and the students' responses made during the solution process provide the evidence that is analyzed by the evidence model. The results of this analysis are data (e.g., scores) that are passed on to the competency model, which in turn updates the claims about relevant competencies. In short, the ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about student competencies via an evidentiary chain (i.e., valid arguments that connect task performance to competency estimates), and are thus valid for their intended purposes. New directions in educational and psychological measurement promote assessment of authentic activities and allow more accurate estimations of students' competencies. Further, new technologies let us administer formative assessments during the learning process, extract ongoing, multi-faceted information from a learner, and react in immediate and helpful ways, as needed.

The following section describes our ideas for embedding assessments within multimedia environments, such as games and simulations.

15.2.2 Stealth Assessment

When embedded assessments are so seamlessly woven into the fabric of the learning environment that they are virtually invisible, we call this stealth assessment (see Shute, Ventura, Bauer, & Zapata-Rivera, in press). Such assessments are intended to support learning, maintain flow, and remove (or seriously reduce) test anxiety, while not sacrificing validity and reliability (Shute, Hansen, & Almond, 2008). In addition, stealth assessment can be accomplished via automated scoring and machine-based reasoning techniques to infer things that are generally too hard for humans (e.g., estimating values of competencies across a network of skills via Bayesian networks).

In learning environments with stealth assessment, the competency model accumulates and represents belief about the targeted aspects of knowledge or skill, expressed as probability distributions for CM variables (Almond & Mislevy, 1999; Shute, Ventura, et al., in press). 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 CM variables (Mislevy, 1994). Task models express situations that can evoke required evidence.

One big question is not about how to collect this rich digital data stream, but rather how to make sense of what can potentially become a deluge of information. Another major question concerns the best way to communicate student-performance information in a way that can be used to easily inform instruction and/or enhance learning. A good solution to the issue of making sense of data, and thereby fostering student learning within immersive environments, is to extend and apply ECD. This provides (a) a way of reasoning about assessment design, and (b) a way of reasoning about student performance in gaming or other learning environments.

We now turn our attention to a literature review and model of a particular key competency – systems thinking skill. Subsequently, we present an example of how to assess this competency within a Quest Atlantis environment (i.e., Taiga Park).

15.2.3 Systems Thinking

> The whole is more than the sum of its parts. (Aristotle)

As noted earlier, rapid changes in today's world have revealed new challenges to and requests from our educational system. Problems facing today's citizens (e.g., global warming, 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 competently in such complex situations requires competence in systems thinking (ST) (Arndt, 2006).

15.2.3.1 Definitions of Systems Thinking

Definitions of systems thinking tend to focus on the relationships between elements in a given environment. Barak and Williams (2007) define ST as the ability to describe and analyze structures and phenomena in natural, artificial, and social environments. Similarly, Salisbury (1996) defines ST as being able to consider all of the elements and relationships that exist in a system, and know how to structure those relationships in more efficient and effective ways. In general, a system can be defined as a group of parts or components working together as a functional unit (Ossimitz, 2000; Salisbury, 1996). A system can be physical, biological, technological, social, symbolic, or it can be composed of more than one of these (Barak & Williams, 2007). Furthermore, many systems are quite complex (e.g., the ecosystem of the world and the human body). To understand the behavior of such complex systems, we must understand not only the behavior of the parts, but also how they act together to form the behavior of the whole. Thus, complex systems are difficult to understand without describing each part and each part must be described in relation to other parts (Bar-Yam, 1997).

Each system consists of closed-loop relations, and system thinkers use diagramming languages and methods to visually represent the relations and feedback structures within the systems. They also use simulations to run and test the dynamics to see what will happen (Richmond, 1993). The National Science Education Standards (National Research Council, 1996) identifies systems as an important and unifying concept that can provide students with a “big picture” of scientific ideas which can then serve as a context for learning scientific concepts and principles. Thus, a strong background in systems thinking is critical to understanding how the world works.

15.2.3.2 Systems Thinking and Its Role in Education

Traditional teacher-centered approaches to education may be less suitable than learner-centered approaches for teaching and bolstering ST skills, especially skills related to considering, understanding, and solving complex problems (Arndt, 2006). This is because in many teacher-centered classrooms students try to assimilate content that is presented by the teacher (Brown, 2003). Students are typically not engaged in ST beyond perhaps repeating back the teacher’s thoughts and interpretations. Although students encounter much content, they do not often learn what to do with it. Thus, this type of learning really does not help much when confronted with novel, complex problems (Arndt, 2006; Richmond & Peterson, 2005). Furthermore, this approach is poorly suited for the transfer of solutions to similar classes of problems. It comes as no surprise that most facts taught and learned via the traditional approach are quickly forgotten (Arndt, 2006). As a consequence, the expectations and needs for a twenty-first century educational system are being inadequately met in settings where students have minimal control of their own learning.

Alternatively, learner-centered approaches are based on the notion that learning is primarily a construction rather than an assimilation process. To learn, the student must construct or reconstruct what is being taken in (Richmond, 1993; Shute, 2007). Students who engage in ST have to actively construct functional relations among relevant components, either mentally or externally.

15.2.3.3 The Competency Model of Systems Thinking

To assess and support ST within a school environment, it is possible to construct indicators for important aspects of systems thinking (Assaraf & Orion, 2005). Having a good competency model should permit educators to collect data about students’ knowledge of and performance on a set of tasks requiring the application of ST skills. This information could then be used to make inferences about students’ current ST competency levels, at various grain sizes, for diagnostic, predictive, and instructional purposes. Our proposed ST competency model consists of three first-level variables: (1) specifying variables and problems in a system, (2) modeling the system, and (3) testing the model via simulation (see Fig. 15.3). Each of these first-level variables has a number of “progeny” and each will now be described in turn.

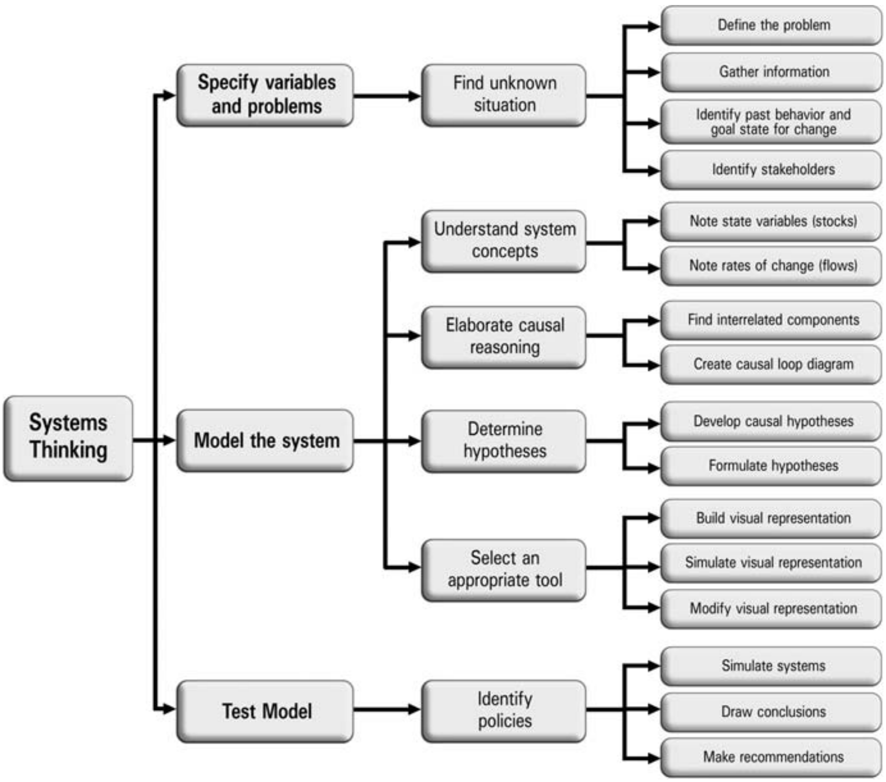


Fig. 15.3 Competency model of systems thinking

Specify Variables and Problems

We believe that the ST process begins by defining problems, formulating and testing potential solutions, and distinguishing fundamental causes of problems (Walker, Greiner, McDonald, & Lyne, 1998). So what exactly is a problem? Jonassen (2004) defines at least two critical features of a problem. The first relates to an unknown entity within some context (i.e., the difference between a goal state and a current state). The second aspect relates to finding or solving the unknown, which must have social, cultural, or intellectual value. Finding the unknown within a problem is important because if no one perceives an unknown, or even a need to determine an unknown, then there is no perceived problem. After defining a problem, system components can be specified in relation to that problem. The best way to determine system components is to answer questions about causality such as: “What causes overpopulation?” Some relevant answers may include: poverty, lack of education, inadequate birth control resources, etc.

Model the System

Conceptual modeling is one of the main tools used to show thinking about a system. The intent of a model is to identify the feedback structures that control behavior. By making these structures explicit, the process helps us share our thoughts with others and simplify complex things. That is, because many elements of a system cannot be observed directly, models help us to visualize and externalize those elements (Jonassen, Strobel, & Gottdenker, 2005; Salisbury, 1996). Fortunately, today's computer technologies allow us to simulate almost any complex situation that we might want to study. Computer simulations also highlight and make visible otherwise hidden processes such as planning, decision making, and evaluation processes (Dörner, 1997). One of the most well-known ST tools is called STELLA (Systems Thinking in an Experiential Learning Laboratory with Animation; see Mills & Zounar, 2001; Salisbury, 1996). Other software applications that are appropriate for creating system diagrams and models in educational settings include: Powersim, Vensim, Modus, Dynasis, and CoLab.

A particularly difficult part of modeling complex systems concerns *interactions* because no action is unilateral in its impact. When one element of a system is changed, it in turn influences other elements of the system. Thus, ST requires an understanding of the dynamic, complex, changing nature of systems (Salisbury, 1996). To illustrate, consider the *butterfly effect* in Chaos Theory, which describes how very small changes, like the flapping of a butterfly's wings in Miami, can affect extremely large systems, like weather patterns in Paris (for more, see Lorenz, 1995). The focus on interactions within ST contrasts with traditional analysis which typically separates the whole into constituent parts (Aronson, 1996). To understand the whole system and its dynamic interactions, the concepts of stocks and flows are crucial (Mills & Zounar, 2001; Sterman, 2000). *Stocks* can be defined as state variables (or accumulations) which hold the current, snapshot state of the system. Stocks completely explain the condition of the system at any point in time and do not change instantaneously. Rather, they change gradually over a period of time. Stocks can represent concrete materials, such as the amount of water in a lake, or abstract concepts, such as level of happiness. *Flows* represent changes, or rates of change. Flows increase or decrease stocks not just once, but at every unit of time (Martin, 1997). For example, the total accumulation of water within a lake is decreased by evaporation and river outlets while it is increased by precipitation and river inlets. Consequently all system changes through time can be represented by using only stocks and flows.

In addition to fully understanding relevant system terms (i.e., stocks and flows, as well as inputs, processes, and outputs), system thinkers must also be concerned with *feedback loops*. Feedback loops are the structures within which all changes occur (Ossimitz, 2000), a closed chain of casual relationships that feeds back on itself (Georgiou, 2007). In other words, feedback represents information about results that supports the system so that the system can modify its work (Salisbury, 1996). The idea of feedback in systems is the most important concept in understanding a problematic situation in a holistic manner, and it also opens the door for quite complex understanding. In interrelated systems we have not only direct, but also indirect effects which may lead to feedback loops. Every action, change in nature, etc. is located within an arrangement of feedback loops.

Feedback loops are represented by causal loop diagrams, and there are two types of feedback: positive (reinforcing) and negative (balancing) (Ossimitz, 2000; Sterman, 2006). Negative feedback intends to achieve some steady state. Positive feedback is self-reinforcing, either in terms of growth (regenerative dynamics) or deterioration (degenerative dynamics). Both growth and deterioration eventually collapse the system in the absence of negative feedback (Georgiou, 2007). World population and birth rate have a positive feedback relationship because large populations cause large numbers of births, and large numbers of births result in a larger population. Each may view the other as a cause (Richmond, 1993), reminiscent of the old chicken-or-egg conundrum. Adding another factor into the equation (e.g., death rate) would be an example of a negative feedback loop influencing population. As a final point on the feedback issue, a proper understanding of feedback loops requires a *dynamic* perspective, in order to see how things appear and then change over time (Ossimitz, 2000).

Another distinction that is made in systems thinking is between open-loop and closed-loop systems. Most people tend to think in a linear manner and use linear thinking (i.e., one cause, one effect) to achieve their goals. Such thinking represents an open-loop system (see Fig. 15.4), where you see a problem, decide on an action, expect a result, and the loop ends (Forrester, 1996).

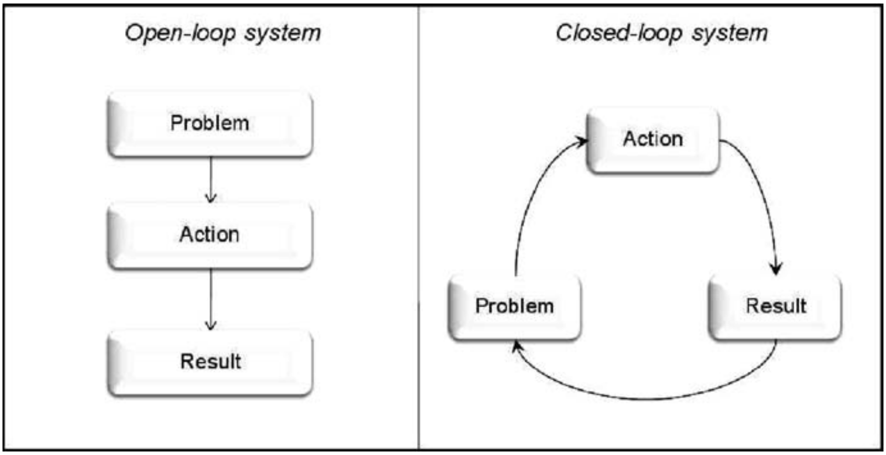


Fig. 15.4 Comparing open-loop and closed-loop systems

However, the real-world does not consist of simple linear relations but of complex relations that are highly interconnected and dynamic. Consequently, the behavior of real systems is often difficult to anticipate because it may be counterintuitive, nonlinear, and irreversible. As a result, linear thinking applied to complex systems is likely to fail (Senge, 1994; Sterman, 2000). To illustrate, think about the factors effecting gasoline prices in the United States. Increasing and decreasing gasoline prices depend on a whole host of factors (e.g., value of the US dollar, supply, demand, OPEC capacity, war effects, Wall Street crises, etc.) and these factors have complex relations with one another. To solve complex problems (like predicting gas prices or tracking hurricane trajectories), people need to think in terms of the "big picture" and about how variables are related to each other rather than in terms of discrete, detailed facts. ST requires knowing about the individual parts of a system, the role each part plays, and how these parts interact to function as a whole (Assaraf & Orion, 2005). In real-life, after gathering information about a problem, this usually leads to some action that produces a result. But in actuality, there is no beginning or end. Instead, the process is iterative (i.e., a closed-loop system; see the right side of Fig. 15.4). So, systems are never totally open. If a system *were* totally open, then it would have no orderly interaction with its environment.

Test the Model

After conceptually modeling the system, the next step involves actually testing out the model. This entails simulating the system (via computational models), running the model, and then drawing conclusions and making decisions based on the obtained results (Richmond & Peterson, 2005). The actual results are compared with the expected results and significant differences must be examined carefully. Differences can be described by computer models. The examination process of unexpected simulation results contains significant opportunities for learning because it requires intensive reflection by the student, as well as adaptation of one's mental model (Sterman, 2000).

15.3 Application of the Stealth Assessment Approach

> Reason does not work instinctively, but requires trial, practice, and instruction in order to gradually progress from one level of insight to another. (Immanuel Kant)

The purpose of this worked example of the systems thinking competency is to test the viability of our stealth assessment approach within an existing immersive game. In the example that follows, we first briefly describe the game (Quest Atlantis: Taiga Park), an immersive, role-playing game set in a modern 3D world (see Barab, Sadler et al., 2007). Next, we present an ECD formulation relating to systems thinking skill as applied and assessed during game play. Finally, we compare a hypothetical player at two different points in time (at the beginning and more advanced stages of learning) in relation to her ST skill.

15.3.1 Quest Atlantis: Taiga Park

Taiga is the name given to 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: (a) the Mulu (indigenous) farmers (e.g., Norbe and Ella); (b) Build-Rite Timber Company (e.g., Manager Lim, Lisa, and Hidalgo); and (c) the K-Fly Fishing Tour Company (e.g., Markeda and Tom). 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. As part of the first mission, a student has to interview 13 different characters throughout the park. Each of them is affiliated with one of the park's main stakeholders. By interviewing the various characters, students "hear" from each one of them about what causes the fish decline in the river – consisting of both opinions and facts about the problem. It soon becomes obvious that the three main stakeholders blame each other, and also that there are more complex problems than just the declining fish problem. At the end of the first mission, students are required to formulate and state an initial hypothesis about the fish-decline problem. This hypothesis is not based on scientific evidence, but on what was heard from the different stakeholders.

For the second mission, students collect water samples from three different sites and analyze the water quality based on six indicators, such as pH level, temperature, and turbidity. Students must submit their interpretation of the water quality data, and also explain which human activities (e.g., fishing, farming, and logging) at each of the three water collection sites cause the problem and how they are interrelated. After completing the second mission, students receive a message from Jesse, Ranger Bartle's intern, which initiates the third mission. The third mission is similar to the second, but focuses on reasoning about the data that has been collected, and drawing a preliminary scientific conclusion based on the hypothesis rendered in the preceding mission.

The fourth mission is set 2 years in the future. It starts with the student being required to name one of the stakeholders as the key culprit in terms of the fish-decline problem. Using a time machine (woven neatly into the narrative), and exploring Taiga 2 years in the future, students can see that ignoring the larger picture (i.e., interrelationships among the stakeholders) and focusing on a simple causal hypothesis and ensuing solution does not work. For instance, suppose that a student blamed the loggers for the fish-decline problem (i.e., logging causes erosion that increases the river's turbidity which leads to gill damage and ultimately death in fish). On the basis of this hypothesis, the park ranger "solves" the problem by ridding the park of the loggers. The future results of the logger-removal decision show that the problem has yet to be solved. Erosion continued because nobody replanted trees, the farmers had to increase farming activities to offset lost revenue from the rent no longer received from the loggers, the fish population continued to suffer and decline, and the park found itself on the brink of disaster. To complete this mission, the student has to explore the future park and explain what has occurred, answering the following questions: (a) Why does blaming just one group create a whole set of different problems? and (b) How can the set of problems be resolved?

The fifth and final mission in Taiga requires students to think of the park as a system, and generate a more coherent hypothesis in relation to the problem, on which the park ranger will act. Students then again employ the time machine to travel 5 years into the future where they view the new version of Taiga Park based on their systemic solution to the problem (i.e., involving both environmentally and economically sustainable solutions). By interviewing different people in Taiga in the future, students identify which changes occurred and how they reflect a socio-scientific solution. In terms of the various *levels* of systems mentioned earlier, students should understand (a) local level systems; i.e., the fragile and interconnected nature of our various ecological systems, like in and around rivers; and (b) socio-economic level systems, like those shown by the entwined relationships among the Taiga stakeholders.

The Taiga Teacher's Guide for this unit notes that activities have been designed around formalized scientific understanding and science learning standards. The five core scientific concepts in the unit include: erosion, eutrophication, water quality indicators (e.g., turbidity, dissolved oxygen), watersheds, and formulating and evaluating hypotheses. Also, through participating in this unit, students are expected to develop valuable skills such as socio-scientific reasoning, scientific inquiry, and scientific decision making. From their experiences in Taiga, students are expected to develop an appreciation for the complexities involved in scientific decision making by balancing ethical, economic, political, and scientific factors (e.g., the best solution from a scientific perspective can be conflicting with political or economic perspectives). Eventually, students are expected to develop deep environmental awareness by appreciating the complexity of environmental problems.

15.3.2 ECD Models Applied to Taiga

Taiga Park, with its requirement for socio-scientific inquiry as well as continuous reflection and revision of current understanding, is an ideal environment to demonstrate the use of ECD for systems thinking. In their role as an expert assistant to the park ranger, students interview stakeholders, collect data, and develop hypotheses about why the fish population in Taiga is declining. Eventually (i.e., in their final mission), the students are expected to recommend a systems-based solution to the park ranger based on their final hypothesis concerning all of the variables affecting the decline in Taiga's fish population.

As described earlier, one important aspect of systems thinking requires a person to conceptualize a model of the system. The main purpose of conceptual modeling is to help a person visualize and externalize elements and relations within a system, and to improve understanding of the dynamic interactions among the different components of a system (i.e., the stocks and flows). To view a problem in a holistic manner, students need to understand how feedback works within a particular system. For instance, feedback loops demonstrate the direct and indirect effects within systems, and causal loop diagrams demonstrate students' understanding of how component changes affect other parts of the system. Once the causal relationships and feedback loops have been established, students should be able to form hypotheses about the relationships within the system. To determine whether a hypothesis is correct, some form of simulation is needed to demonstrate the stated relationships between system components. This process enables students to then modify the original hypothesis. Fortunately, in Taiga Park, there is a time machine. This clever narrative device permits one to simulate consequences of particular actions at various points of time in the future.

Figure 15.5 shows a conceptualization of the ECD models for a fragment of the ST competency (i.e., *Model the System*). Notice that "competency model" and "evidence model" are the same terms as we used in the previous ECD discussion. However, when extending to game environments, we use the term "action model" instead of task model. An action model reflects the fact that we are dynamically modeling students' actions within the particular game. These actions form the basis for gathering evidence and rendering inferences and may be compared to simpler

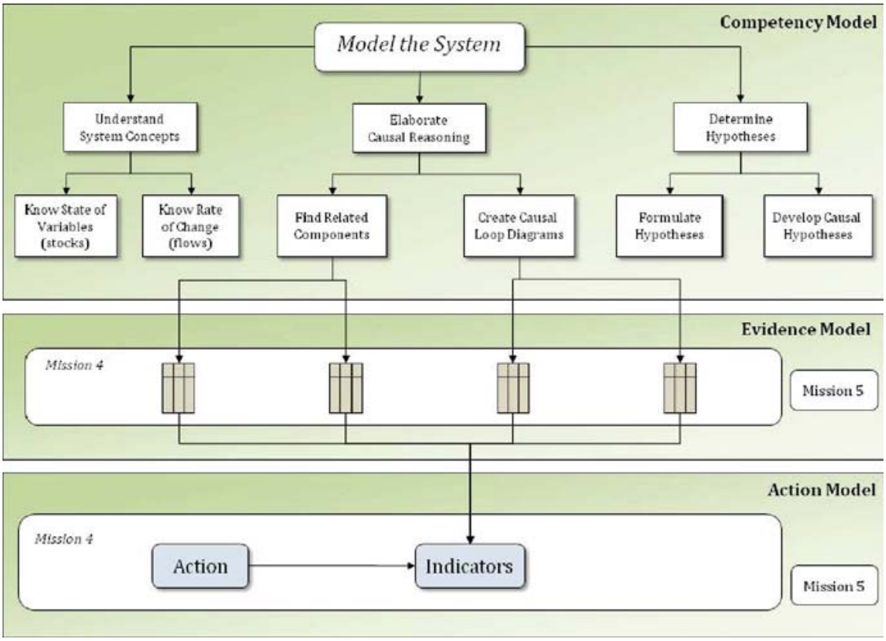


Fig. 15.5 Conceptualization of ECD models applied to Taiga

task responses as with typical assessments. The lined boxes shown within the evidence model denote what are called conditional probability tables (CPTs). These CPTs represent the statistical relations (or “glue”) between the indicators (observable) and competencies (unobservable). Finally, note that “mission” is used to define a set of required actions within a particular quest.

Competency Model: By the time students reach Mission 4 in Taiga, they have (a) interviewed a variety of people who have a stake in the park, (b) collected water samples from three different points along the river, and (c) taken snapshots at five observation posts located along the river. Thus in mission 4, students need to demonstrate an understanding of how the water quality indicators (e.g., turbidity, pH level, temperature) relate to the activities along the river – specifically in relation to their effects on the fish population. Additionally, students should be able to draw a causal diagram that shows the stocks and flows of the components that are reducing the population of fish in the river.

Evidence Model: This model is established to determine how the observable aspects of the students’ actions in the game may be used (i.e., collected and aggregated) as evidence for the competency variables. The evidence model contains: (a) outcomes from the assigned tasks such as diagrams created or short answers provided to specific questions, (b) rules for scoring the student submissions, and (c) indicator weights in relation to associated competencies.

Action Model: Similar to the task model, 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. Some of the required actions are sequential in nature and must be completed in order to proceed within the mission. Other actions can occur at any point in time, and as often as desired. Table 15.1 lists a few representative actions and their indicators relevant to various Taiga missions.

In the current version of Taiga, students write and submit short essays to their teachers as a required part of the missions. The teacher then reviews the essays, using a set of rubrics to score them. For example, a student may receive maximum points (and earn a badge) for an essay answer that demonstrates: (a) an ability to interpret water-quality indicators, (b) an understanding of ecological processes, and (c) the capability to integrate evidence (obtained during missions) and the associated processes. Students falling short of the criteria are advised to visit the water expert at Taiga to discuss the water indicators and ecological processes again. They are also told to revise and resubmit their essays if they wish to receive the badge of completion.

In addition to the essays, students can create and submit *causal loop diagrams* (demonstrating the stocks and flows within the system and their cause-effect relationships). In the current version of the game, such diagrams may be uploaded as an attachment to student essays, but they are optional. One problem with the current implementation is the large burden it places on teachers to not only monitor their students’ game play, but additionally to carefully read and score all essays, interpret and assess the quality of all submitted causal diagrams, as well as provide feedback to support students’ learning. Also, there may be ambiguity in diagrams and

Table 15.1 List of actions and associated indicators

Action	Indicators
Summarize water quality indicators along the river	Accurately note water quality indicators for 3 points along the river Accurately note whether indicators signify good or bad water quality
Explain how water-quality data account for fish death	Correctly explain how the indicators are symptoms of erosion and eutrophication Correctly link these ecological processes to the population of fish in Taiga River
Explain how the various stakeholders contribute to the fish-decline problem	Correctly identify stakeholders and their main activities near the river Correctly relate these activities to erosion and eutrophication
Create causal loop diagram	Include complete set of variables and links in the diagram Accurately identify relationships among variables (positive or negative)
Evaluate a hypothesis	Correctly identify one group responsible for the problem at Taiga Accurately explain and/or depict how this group's activities lead to ecological processes detrimental to the fish

subjectivity in assessing, on the teachers' parts. Moreover, crafting causal diagrams, we believe, should be an integral (not optional) part of the game.

15.3.2.1 Tools to Automatically Assess Causal Diagrams

If causal diagrams were required in the game, how could we automate their assessment? Solving this issue would reduce teachers' workload, increase the reliability of the scores, and clearly depict students' current mental models (or conceptualizations) of various systems operating within Taiga. Students' causal diagrams can be created using one of several computer-based tools designed for this purpose (e.g., CmapTools, by Cañas et al., 2004; freeware which can be downloaded from: http://cmap.ihmc.us/conceptmap.html). There are currently quite a few tools and technologies emerging whose goal is to externalize and assess what are otherwise internal conceptions (e.g., see Shute, Jeong, Spector, Seel, & Johnson, in press). The tool that we focus on in this illustration is an Excel-based software application called jMap (Jeong, 2008; Shute, Jeong, & Zapata-Rivera, in press), 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 toward an expert's or the aggregated group model (for more information about the program, including links and papers, see: http://garnet.fsu.edu/∼ajeong).

With jMap, students create their causal maps using Excel's autoshape tools. Causal links are used to connect a collection of variables together, and link strength may be designated by varying the thicknesses of the links (not relevant in the following worked example). In jMap, comparisons between a student's and a target map[1] begin by automatically coding/translating each map into a transitional frequency matrix. For instance, if the target map contained eight variables comprising a complete causal diagram, this would translate to an 8×8 frequency matrix representing all pairwise linkages (see Table 15.2). Each observed link within the student's map is recorded into the corresponding cell of the matrix.

Table 15.2 Example of a transitional frequency matrix

Transitional Frequency Matrix	Taiga Park income	Need more logging	Cutting trees	Soil erosion	Sediment in water	Temperature of water	Dissolved oxygen	Fish population
Taiga Park income								
Need more logging								
Cutting trees								
Soil erosion								
Sediment in water								
Temperature of water								
Dissolved oxygen								
Fish population								

Once all (i.e., student and expert maps) have been automatically tabulated into transitional frequency matrices, jMap can be used to superimpose: (a) the map of one learner produced at one point in time over a map produced by the same learner at a later point in time; (b) the map of one learner over the map of a different learner; or (c) the map of a learner over the map of an expert. jMap can also be used to aggregate all the frequencies across the frequency matrices of multiple learners to produce an aggregate frequency matrix representing the collective group. As a result, the resulting collective group map can also be superimposed over an individual learner's map or an expert map. Users (e.g., teachers, researchers, students,

---

[1] The target map is usually an expert's map, but may be another student map (e.g., the same student at different times, a different student, or even a group of students). See Shute, Jeong, and Zapata-Rivera (in press) for examples.

etc.) can toggle between maps produced over different times to animate and visually assess how maps change over time and see the extent to which the changes are converging toward an expert or group map. Additional jMap tools enable users to compile raw scores to compare quantitative measures (e.g., the percentage of shared links between the compared maps).

In this proposed scenario, and as part of their gaming mission, students would draw their causal diagrams using jMap, which would contain a collection of relevant system concepts or stocks. Students would choose relevant variables from the collection, and link them together, similar to completing a puzzle, into a causal diagram. This activity would (a) take place within the Taiga narrative (e.g., as part of a task assigned to the student by Ranger Bartle), and (b) demonstrate students' emerging understanding of the interrelatedness of relevant concepts. The submitted maps would then be automatically compared in terms of propositional structure with an expert (or target) map. Higher similarity indices between the two would lead to higher estimates for the relevant competency.

15.3.2.2 Adding Stealth Assessment to Taiga

To illustrate this automated, evidence-based assessment methodology within Taiga, we implemented a part of the ECD model relating to systems thinking skill, and focused on the competency: *Model the System*.

Figure 15.6 shows the initial state of the network. When a student performs an action in the game (e.g., creates a causal loop diagram), relevant indicators are calculated. For this example, the indicators include (a) accuracy/completeness of the variables included in the diagram, and (b) accuracy of the links established (i.e., positive versus negative relations). These comprise the set of indicators associated with that particular node (see Table 15.1). The indicator data, derived from the jMap tool, are then automatically inserted into the Bayes net which is instantly updated with new probability values propagated throughout the network.

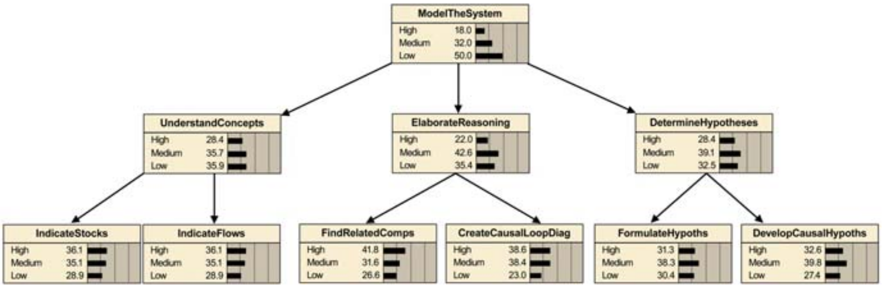


Fig. 15.6 Initial Bayesian model for a fragment of systems thinking skill

Consider a hypothetical student named Clara. Suppose we have two causal loop diagrams obtained from her at two different points in time: during an early mission in Taiga, and then during her final mission. During the early mission, Clara blamed

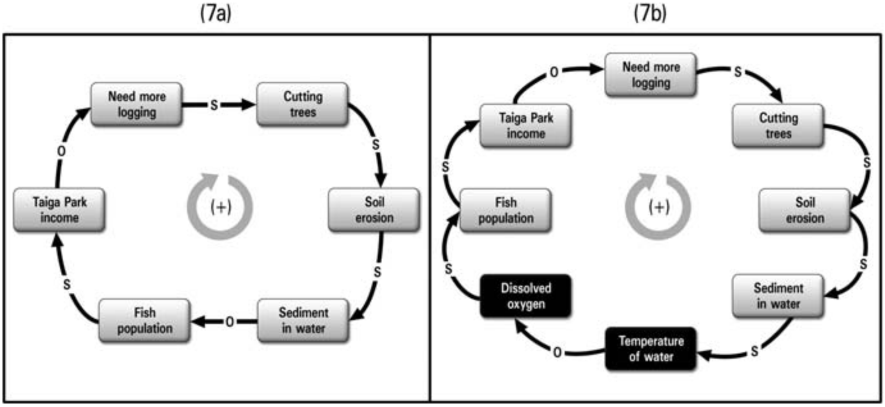


Fig. 15.7 Clara's causal loop diagram at Time 1 (**a**) and an expert diagram of the system (**b**)

the decline-in-fish-population problem solely on the loggers. Her causal loop diagram at that point is shown in Fig. 15.7a (see left panel). The full set of variables available in the jMap collection includes those shown in her diagram, as well as others such as dissolved oxygen in the water, temperature of the water, pH level of the water, and so on. The relationships between variables are also recorded directly in the diagram using an "S" (for same, denoting a positive function) or an "O" (for opposite, for an inverse function).

At this relatively early stage of learning, Clara appears to have a basic understanding of what is going on in the river relative to the logging business, but does not yet fully understand all of the variables that cause a decrease in the fish population. If her diagram was compared to an expert's (using jMap), her errors of omission would suggest that she believes sediment in the water directly and negatively affects the fish population. However, sediment in the water actually serves to increase water temperature, which in turn causes a decrease in the dissolved oxygen. Inadequate oxygen would cause fish to die. This provides the basis for valuable feedback to Clara, which could be automatically generated, or provided by the teacher (e.g., "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 is the reason the fish are dying – they do not have enough oxygen"). In addition, the lab technician (or another knowledgeable character in Taiga) could provide feedback in the form of a causal loop diagram, explicitly including those variables in the picture. That way, she can see for herself what she had left out. See the right panel in Fig. 15.7b for an example of an expert diagram, highlighting her omitted variables and links.

When she visits Taiga 2 years in the future, Clara would quickly realize that her simple conceptualization of the problem (i.e., blaming just a single group of Taiga stakeholders – the loggers) and the ensuing solution (i.e., Ranger Bartle's banning the loggers from Taiga Park) was in vain. That is, 2 years into the future, she sees converging evidence that the fish population is still suffering – perhaps even worse than before. Over the course of additional actions and interactions in Taiga (e.g., comparing photos taken along the river at different times, interviewing people in the present and the same people again in the future), she gradually understands the ramifications of her previous solution. That is, because the loggers are gone, the Mulu farmers had to increase their farming operations to offset their lost income (from loggers' rent money). This increase in farming operations resulted in more nutrients from fertilizer running off into the river and affecting the ecosystem (negatively for the fish – positively for the algae); and more toxic waste running off into the river from increased use of pesticides. Many actions and interactions later, Clara eventually comprehends the functional relationships among all three stakeholders and sees how they all are to blame for the problem. This holistic (system) understanding can now provide the basis for an effective solution to the declining-fish-population problem that concurrently addresses all aspects of the issue (i.e., the effects of farming, logging, and fishing tournaments on the fish population). Consequently, she draws a more comprehensive causal diagram (see Fig. 15.8) and recommends various regulations on all three stakeholders to Ranger Bartle.

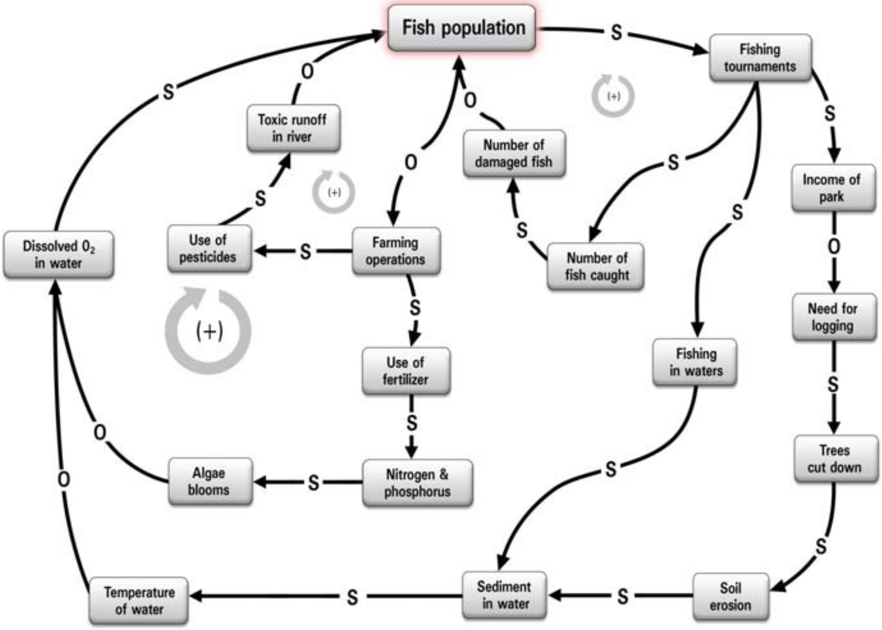
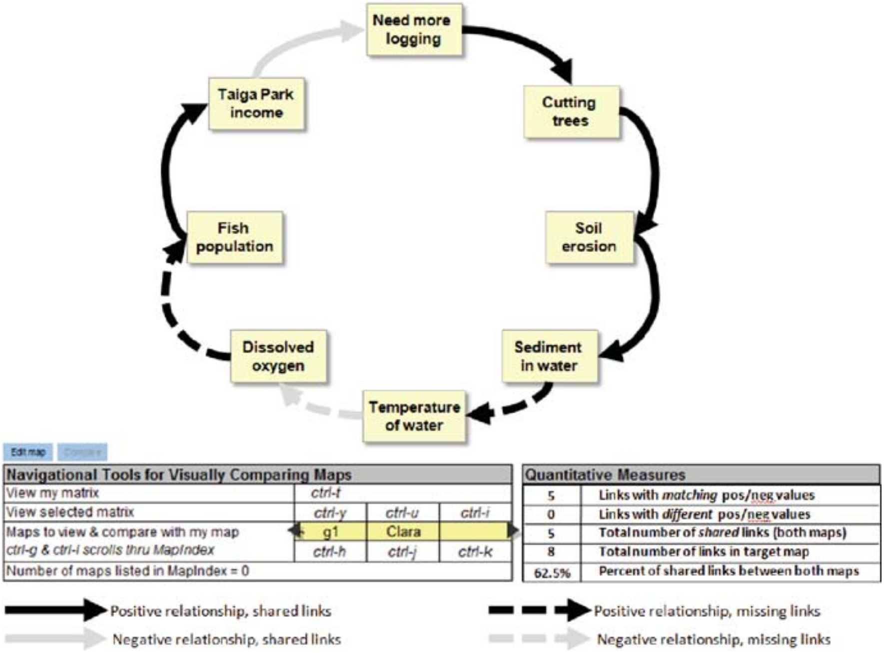


Fig. 15.8 Clara's causal loop diagram – Time 2

So how does jMap derive indicator values to feed into the Bayes net? Let us look at the jMap analysis comparing Clara's Time 1 map to an expert map. As shown earlier, Clara demonstrated incomplete modeling of the system based on her performance on relevant indicators. A screen capture from jMap is shown in Fig. 15.9. Here, jMap's generated diagram uses colored links to clearly and visually identify



Navigational Tools for Visually Comparing Maps			
View my matrix	*ctrl-t*		
View selected matrix	*ctrl-y*	*ctrl-u*	*ctrl-i*
Maps to view & compare with my map	g1	Clara	
ctrl-g & ctrl-l scrolls thru MapIndex	*ctrl-h*	*ctrl-j*	*ctrl-k*
Number of maps listed in MapIndex = 0			

Quantitative Measures	
5	Links with *matching* pos/neg values
0	Links with *different* pos/neg values
5	Total number of *shared* links (both maps)
8	Total number of links in target map
62.5%	Percent of shared links between both maps

Fig. 15.9 jMAP interface showing a Clara's Time 1 map overlaid on the expert's map

differences between two selected maps – in this case between Clara's Time 1 map and the expert map. Dashed arrows denote *missing links* (i.e., links that are present in the expert map but missing in the student map), and solid arrows denote shared links, which match in terms of identical positive/negative assigned values. The color black represents positive relations and grey represents negative ones. jMap also has the option to represent link strengths (e.g., weak, medium, and strong influences), but we are ignoring link strength in this scenario to make the example easier to understand. By visual inspection, we can see that Clara has omitted three links (and two important variables) in her causal loop diagram relative to the expert's map (shown by the three dashed arrows).

In addition to the standardized maps, the jMAP interface includes two tables, as shown below the map in Fig. 15.9. The table on the left includes navigational tools. These allow the user (e.g., teacher, student, researcher) to easily move among all possible maps using control-key functions, showing the map, the matrix, or both, and compared to the expert model or another model, such as a group model. The table on the right labeled "Quantitative Measures" provides an indication of the similarity between the current map (in this case, Clara at Time 1) and the expert map. The percentage of shared links between the two maps is 62.5%.

If cut-off values were assigned (e.g., 0–33% = low; 34–66% = medium; 67–100% = high), then Clara's accuracy/completeness of her diagram would be classified as medium. Furthermore, because she had created the correct relations of

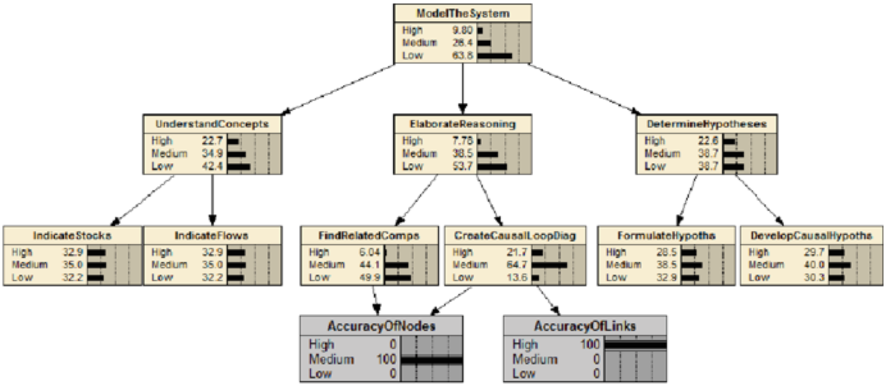


Fig. 15.10 Bayesian model for Clara at Time 1

the links in her diagram (i.e., positive versus negative functions), she would receive a score of "high" on that indicator. These indicator outcomes are then inserted into the Bayes net (see Fig. 15.10).

Once the information is inserted into the Bayes net, it is propagated throughout the network to all of the nodes, whose estimates are subsequently altered. For instance, her Time 1 estimate for the competency, "Create causal loop diagram" is medium; her "elaborate reasoning" competency, however, is estimated at low, as is her overall competency, "model the system." She has more work to do in Taiga, and this analysis and diagnosis targets particular areas for improvement.

By the final Taiga mission, as evidenced in her causal loop diagram shown in Fig. 15.8, Clara has acquired a good understanding of the various systems in Taiga. Her final causal diagram shows the interwoven processes of erosion and eutrophication taking place along the river from the three Taiga communities. The Bayesian model of Clara at Time 2 (not shown) provides evidence of her ability to understand the relationships among system components, with an overall estimate of her "model the system" competency to likely be "high" (i.e., $p(\text{high}) = 0.60$; $p(\text{medium}) = 0.36$; and $p(\text{low}) = 0.04$). This example shows how the outcomes of actions carried out within the game can be used to infer different levels for important competencies in a game environment.

15.4 Summary and Discussion

We presented an innovative approach for embedding evidence-based assessment within an immersive game environment to estimate students' evolving system thinking skills. The ongoing assessment information is intended to provide the basis for bolstering students' competency levels within the game, directly and indirectly. Our approach represents an extension of ECD, which normally entails assessment tasks (or games, simulations, etc.) being developed at the end of the ECD process. But in this chapter, we illustrated how we can employ an evidence-based approach using an existing game.

The steps of this approach involve the following: (a) define the competency model for systems thinking, independently from the game, via an extensive literature review which is validated by experts (the validation is currently underway); (b) determine indicators of the low-level nodes in the CM relative to particular game actions; (c) specify scoring rules for the indicators; and (d) develop evidence models that statistically link the indicators to particular nodes in the CM via Bayes nets (or any other method for accumulating evidence). Our hypothesis is that the CM (stripped of specific "indicators") should be transferable across environments that require students to engage in systems thinking skill. This type of "plug and play" capability would make the CM scalable, which comprises part of our plans for future research. Finally, we presented just one example of automatically assessing a component of ST (i.e., creating causal loop diagrams). However, other nodes in the model can be easily and automatically assessed, like those that relate to acquiring relevant knowledge (e.g., water-quality indices like turbidity and alkalinity) and skill at gathering pertinent information in the environment (e.g., collecting water samples from different parts of the river and making sense of the data). Additional attributes (e.g., teamwork and communication skills) can similarly be assessed in the game, providing that a CM has been developed and indicators fully identified.

Another near-future research plan includes examining our stealth assessment approach under conditions where there are multiple, valid solutions to a problem (i.e., less-structured scenarios compared to Taiga Park). For instance, we are currently exploring and analyzing other worlds in Quest Atlantis and deriving assessments that pertain to (a) creative problem solving, and (b) multiple-perspective taking, both identified as key competencies for the twenty-first century. In less-structured environments, multiple solutions can be identified by experts in the content area, and each possible solution then converted to a Bayesian network. The higher level competency nodes (reflecting mastery of rules applicable to a wide range of problems within a content area) should be similar, while the lower level indicators reflect different approaches to problem solving (Conati, 2002).

The main problem that we seek to address with this research is that educational systems (in the US and around the world) are facing enormous challenges that require bold and creative solutions to prepare our students for success in the twenty-first century. Part of the solution will require a strong focus on students developing the ability to solve complex problems in innovative ways, as well as the ability to think clearly about systems. We need to identify ways to fully engage students through learning environments that meet their needs and interests (e.g., through well-designed educational games). When coupled with online collaboration with other students (locally and from around the world), such environments additionally have the potential to develop students' communication skills and creative abilities as they become exposed to diverse cultures and viewpoints.

We maintain that not only is it important to determine the skills needed to succeed in the twenty-first century, but also to identify particular methods for designing and developing assessments that are valid and reliable and can help us meet the educational challenges confronting us today. One looming challenge, as mentioned earlier, concerns the need to increase student engagement. Thus, we have chosen to embed our stealth assessment approach and associated tools within the context of an immersive game (e.g., Quest Atlantis). Through such games, learning takes place within complex, realistic, and relevant environments (although even fantasy games, such as quests within legendary kingdoms involving nonhuman characters, can be used as the basis for assessment and support of valuable skills). Moreover, games can provide for social negotiation where students learn to communicate and collaborate with others on team quests. Such skills are integral parts of many games, and are crucial for players to complete missions. This design feature can help students consider and respect multiple perspectives from other team members who play different roles and have different strengths and backgrounds. Games can also engender ownership of learning since students can choose to complete a particular quest or explore less well-trodden paths to satisfy their curiosity.

The challenge for educators who want to employ 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). Our 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 permits 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/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 particular sequences of actions (e.g., as in Quest Atlantis) 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.

Currently, Quest Atlantis employs a system that enables teachers to view their students' progress during their missions via a web-based Teachers Toolkit panel. This enables teachers to 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 simply 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. This harnesses the power of formative assessment to support learning.

In conclusion, our proposed solution using ECD, stealth assessment, and automated data collection and analysis tools is meant to not only collect valid evidence of students' competency states, but to also reduce teachers' workload in relation to managing the students' work (or actually "play") products. This would allow teachers, then, to focus their energies on the business of fostering student learning. If the game was easy to employ and provided integrated and automated assessment tools as described herein, then teachers would more likely want to utilize the game to support student learning across a range of educationally valuable skills. Our proposed ideas and tools within this worked example are intended to help teachers facilitate learning, in a fun and engaging manner, of educationally valuable skills not currently supported in school. Our future research plans include implementing a full systems thinking stealth assessment into the Taiga Park virtual world to test its efficacy in support of students as well as teachers.

References

Abedi, J., & O'Neil, H. F. (2005). Assessment of noncognitive influences on learning. *Educational Assessment*, *10*, 147–151.

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

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

Aronson, D. (1996). *Overview of system thinking*. Retrieved January 8, 2009, from http://www.thinking.net/Systems_Thinking/OverviewSTarticle.pdf

Assaraf, O. B.-Z., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, *42*(5), 518–560.

Barab, S. A. (2006, Winter). 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.

Barak, M., & Williams, P. (2007). Learning elemental structures and dynamic processes in technological systems: A cognitive framework. *International Journal of Technology and Design Education, 17*(3), 323–340.

Bar-Yam, Y. (1997). *Dynamics of complex systems. Studies in nonlinearity*. Reading, MA: Addison-Wesley.

Brown, K. L. (2003). From teacher-centered to learner-centered curriculum: Improving learning in diverse classrooms. *Education*, *124*(1), 49–54.

Cañas, A. J., Hill, R., Carff, R., Suri, N., Lott, J., Eskridge, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), *Concept maps: Theory, methodology, technology*. Proceedings of the First International Conference on Concept Mapping (pp. 125–133). Pamplona: Universidad Pública de Navarra.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence*, *16*(7–8), 555–575.

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

Dörner, D. (1997). *The logic of failure: Recognizing and avoiding error in complex situations.* New York: Metropolitan Books.

Farkas, G. (2003). Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology*, *29*, 541–562.

Forrester, J. W. (1996). *System dynamics and K-12 teachers*. Retrieved August 8, 2008, from Massachusetts Institute of Technology (MIT), Systems Dynamics in Education Project Web site: http://sysdyn.clexchange.org/sdep/Roadmaps/RM1/D-4665-4.pdf

Gee, J. P. (2004a). *Situated language and learning: A critique of critical schooling.* London: Routledge.

Gee, J. P. (2004b). *What video games have to teach us about literacy and learning.* London: Palgrave Macmillan.

Georgiou, I. (2007). *Thinking through systems thinking.* London: Routledge.

Jeong, J. C. (2008). *Discussion analysis tool (DAT).* Retrieved December 22, 2008, from http://garnet.fsu.edu/∼ajeong/DAT

Jonassen, D. H. (2004). *Learning to solve problems: An instructional design guide*. San Francisco, CA: Pfeiffer.

Jonassen, D., Strobel, J., & Gottdenker, J. (2005). Model building for conceptual change. *Interactive Learning Environments*, *13*(1), 15–37.

Lorenz, E. N. (1995). *The essence of chaos*. Seattle, WA: University of Washington Press.

Martin, L. A. (1997). *Road map 2: Beginner modeling exercise*. MIT System Dynamics in Education Project. Retrieved August 5, 2008, from Massachusetts Institute of Technology (MIT), Systems Dynamics in Education Project Website: http://sysdyn.clexchange.org/sdep/Roadmaps/RM2/D-4347-7.pdf

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

Mills, I. J., & Zounar, E. D. (2001). *On the application of system dynamics to the integration of national research and K-12 education*. Paper presented at the International Conference on Engineering Education, Oslo & Bergen, Norway.

Mislevy, R. J. (1994). Evidence and inference in educational assessment, *Psychometrika*, *12*, 341–369.

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

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

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Ossimitz, G. (2000). *The development of systems thinking skills using system dynamics modeling tools*. Retrieved August 13, 2008, from Universität Klagenfurt, Institut für Mathematik, Statistik und Didaktik der Mathematik Website: http://wwwu.uni-klu.ac.at/gossimit/sdyn/gdm_eng.htm

Park, O., & Lee, J. (2003). Adaptive instructional systems. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 651–685). Mahwah, NJ: Lawrence Erlbaum.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

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

Richmond, B., & Peterson, S. (2005). *An introduction to systems thinking: STELLA software*. Lebanon, NH: High Performance Systems, Inc.

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.

Salisbury, D. F. (1996). *Five technologies for educational change: Systems thinking, systems design, quality science, change management, instructional technology.* Englewood Cliffs, NJ: Technology Publications.

Senge, P. M. (1994). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday/Currency.

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). New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

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

Shute, V. J., Dennen, V. P., Kim, Y. J., Donmez, O., & Wang, C.-Y. (2008). *21st century assessment to promote 21st century learning: The benefits of blinking.* Unpublished manuscript, Florida State University, Tallahassee.

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., Spector, J. M., Seel, N. M., & Johnson, T. E. (2009) Model-based methods for assessment, learning, and instruction: Innovative educational technology at Florida State University. In M. Orey (Ed.), *2009 Educational media and technology yearbook*, Westport, CT: Greenwood Publishing Group.

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, & J. M. Spector (Eds.), *Instructional design for complex learning*. New York: Springer.

Shute, V. J., Lajoie, S. P., & Gluck, K. A. (2000). Individualized and group approaches to training. In S. Tobias & J. D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military* (pp. 171–207). New York: Macmillan.

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. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: Theories and applications*. Philadelphia: Routledge/LEA.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). The MIT Press.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence* (pp. 3–37). New York: Cambridge University Press.

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.

Sterman, J. (2000). *Business dynamics: Systems thinking and modeling for a complex world.* Boston: Irwin/McGraw-Hill.

Sterman, J. (2006). Learning from evidence in a complex world. *American Journal of Public Health*, *96*(3), 505–514.

Walker, P. A., Greiner, R., McDonald, D., & Lyne, V. (1998). The tourism futures simulator: A systems thinking approach. *Environmental Modeling and Software*, *14*(1), 59–67.

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