Tech., Inst., Cognition and Learning, Vol. 8, pp. 137–161
Reprints available directly from the publisher
Photocopying permitted by license only

Conceptual Framework for Modeling, Assessing and Supporting Competencies within Game Environments

VALERIE J. SHUTE*, ISKANDARIA MASDUKI AND OKTAY DONMEZ

Florida State University, 3205G Stone Building,
PO Box 3064453, Tallahassee, FL 32306

The first challenge of accomplishing the goals of any successful instructional system depends on accurately assessing learners and leveraging the information to improve learning (e.g., Conati, 2002; Park & Lee, 2003; Shute, Lajoie, & Gluck, 2000; Snow, 1994). This paper describes an approach for modeling key competencies and developing valid assessments embedded within an immersive game. Specifically, we describe theoretically-based research relating to stealth assessment, diagnosis, and instructional decisions, operational within an immersive game environment. Stealth assessment and diagnosis occur during the learning (playing) process, and instructional decisions are based on inferences of learners' current and projected competency states.

Keywords: Bayesian networks, evidence-centered design, games, stealth assessment, systems thinking

Can games be used to support meaningful learning? Most likely the answer is yes, conditional on more research being conducted in this area. In general, we believe that (a) 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); and (b) learning environments should thus be interactive, provide ongoing feedback, grab and sustain attention, and have appropriate and adaptive levels of challenge—i.e., the features of good games (e.g., Prensky, 2001; Salen & Zimmerman, 2004).

*Corresponding author: vshute@fsu.edu

Along the same lines, Gee (2003) has argued that the secret of a good game is not its 3D graphics and other bells and whistles, but its underlying architecture where 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. More recent reports (e.g., Shute, Rieber, & Van Eck, in press; Thai, Lowenstein, Ching, & Rejeski, 2009) contend that well-designed games can act as transformative digital learning tools to support the development of skills across a range of critical educational areas. In short—well designed games have the potential to support meaningful learning across a variety of content areas and domains.

A common challenge in both learning and gaming environments relates to the provision of formative feedback. It is an important component of learning (Shute, 2008) and also a critical part of good game design where players are provided with challenges that are commensurate with their skill level, and given feedback to let them know how they are progressing. Ideally learners or players are assessed and provided with feedback in a natural and seamless manner that supports learning while not disrupting the fun of game play.

Purpose

This paper describes a conceptual framework and tools for modeling, assessing, and supporting important competencies via assessments embedded within immersive games. Our modeling efforts extend an existing evidence-centered design (ECD) approach formulated by Mislevy, Steinberg, and Almond (2003) and employ Bayesian networks (Pearl, 1988). Inferences – both diagnostic and predictive – about current competency states are handled by Bayes nets and used directly in the student models to handle uncertainty. To make these ideas more concrete, we focus on an existing 3D immersive game called Quest Atlantis: *Taiga Park* (e.g., Barab & Jackson, 2006; Barab et al., 2007a; Barab et al., 2007b), and demonstrate how evidence is gathered and interpreted in relation to one of our targeted competencies: systems thinking skill.

To accomplish our goal of developing really good assessments embedded in games 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. Afterwards, we present a brief literature review and comprehensive model associated with the systems thinking competency, as well as a description of how

these ideas actually play out within an existing immersive game – Quest Atlantis: Taiga Park.

ASSESSMENT METHODOLOGY: EVIDENCE-CENTERED DESIGN

The fundamental ideas underlying ECD came from Messick (1994). This process begins by identifying what should be assessed in terms of knowledge, skills, or other learner attributes. These variables cannot be observed directly, so behaviors and performances that demonstrate these variables need to 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).

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 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 requires an assessor to: (a) define the claims to be made about students' competencies, (b) establish what constitutes valid evidence of the claims, and (c) determine the nature and form of tasks or situations that will elicit that evidence. These three actions map directly onto the three main models of ECD shown in Figure 1.

A good assessment has to elicit behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence

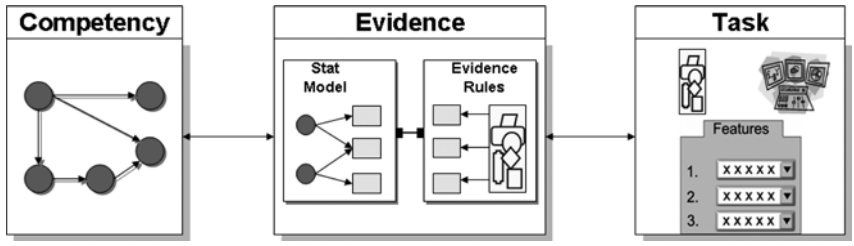

FIGURE 1
Three main models of an evidence-centered assessment design.

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 heart of assessment design.

Competency Model

What collection of knowledge, skills, and other attributes should be assessed? 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 contexts.

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. 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.

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., 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).

Design and Diagnosis

As shown in Figure 1, assessment design flows from left to right, although in practice it's more iterative (i.e. the assessor identifies competency variables, establishes the criteria for evaluating performance, and designs/administers the tasks to elicit performance evidence). 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. Next we describe our stealth assessment idea.

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, in press; Shute, Ventura, et al., 2009). Such assessments are intended to support learning, maintain flow, and remove (or 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., 2009). 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. In short, ECD 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 brief literature review and model of a particular competency: systems thinking skill. Subsequently, we present an example of how to assess this competency within an immersive game.

Systems Thinking

The whole is more than the sum of its parts. ~ Aristotle

Rapid changes in today's world have revealed new challenges to and requests from our educational system. Problems facing today's citizens (e.g., massive oil spill in the Gulf of Mexico, 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 sub-systems 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).

Definition of ST

The systems thinking construct refers to one's ability to understand the relationships between elements in a given environment. 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).

Our ST Competency Model

To assess and support ST within a learning environment, it's 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 activities 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 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 Figure 2). Each of these first-level variables will now be described in turn.

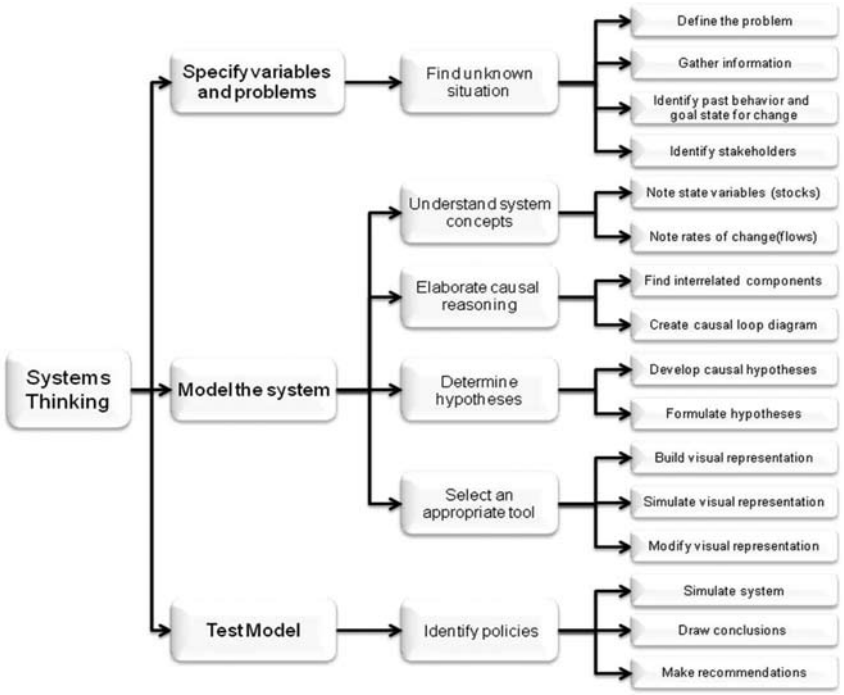

FIGURE 2
Competency model of systems thinking.

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). 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 involves explicating important variables and their relationships relative to a particular system. A variety of tools exist to support conceptual modeling, and the intent of a model is to identify the feedback structures that control behavior. Because many elements of a system can't be observed directly, models help us to visualize and externalize those elements (Jonassen,

Strobel, & Gottdenker, 2005; Salisbury, 1996). 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 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 over a period of time (such as the amount of water in a lake, or abstract concepts like the 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 understanding system concepts (i.e., stocks and flows, as well as inputs, processes, and outputs), system thinkers must also be concerned with *feedback loops*. Feedback represents information about results that supports the system so that the system can modify its work (Salisbury, 1996). The idea of feedback 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, represented by causal loop diagrams.[2]

Another distinction that's made in systems thinking is between open- vs. closed-loop systems. Most people think in a linear manner (i.e., one cause, one effect) to achieve their goals. Such thinking represents an open-loop system (see Figure 3), where you see a problem, decide on an action, expect/observe a result, and the loop ends (Forrester, 1996). 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).

[2] System dynamics uses two ways to represent the dynamic structures: causal loop diagrams and stock and flow diagrams. Causal loop diagrams are visualizations representing structural interrelationships between the parts of a system. They show the process of the system with arrows going from one element to another and back again (Senge et al., 2000), but these diagrams do not explain the unique qualities of a particular situation. Stock and flow diagrams represent the structure of a system with more detailed information than causal loop diagrams (Lane, 2008).

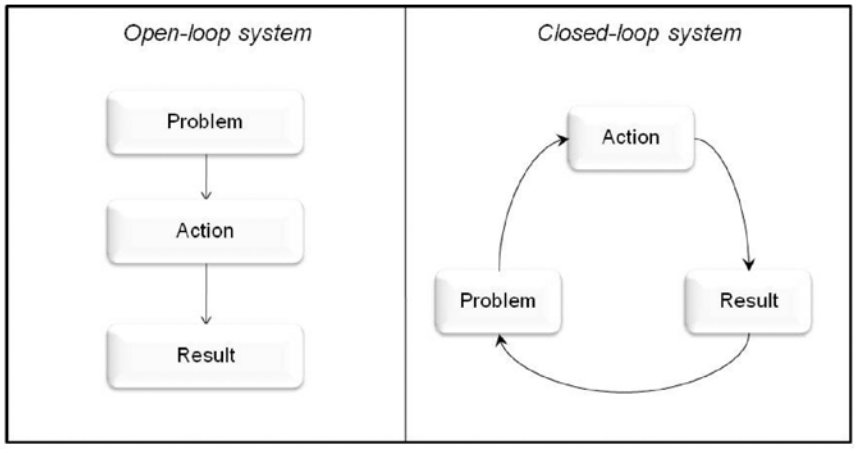

FIGURE 3
Comparing open-loop and closed-loop systems.

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).

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. The examination process of unexpected simulation results contains excellent opportunities for learning because it requires intensive reflection by the student, as well as adaptation of one's mental model (for more, see Jensen & Brehmer, 2003; Sterman, 2000).

APPLICATION OF THE STEALTH ASSESSMENT APPROACH

The purpose of our example in this section is to test the feasibility of our stealth assessment approach within an existing immersive game. In the example that follows, we first describe the game (Quest Atlantis: Taiga Park), an immersive, role-playing game set in a modern 3D world (see Barab et al, 2007b). Next, we present an ECD formulation relating to systems thinking skill as applied to

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.

Quest Atlantis: Taiga Park

Taiga is the name given to a beautiful virtual park with a river running through it (Barab et al., 2007a; Zuiker, 2007). The park is populated by several groups of people who use or depend on the river in some capacity. In addition to the park ranger (Ranger Bartle), the three stakeholders include: (a) the Mulu (indigenous) farmers; (b) Build-Rite Timber Company; and (c) 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 focuses on the declining fish population in the Taiga River. Students participate in this world by helping Ranger Bartle figure out how to solve the dying-fish problem and thus save the park.

As part of the first mission, a student has to interview thirteen different characters throughout the park and "hear" from each one of them about what is causing the fish to die. The interviewees' input consists of both opinions and facts about the problem. It soon becomes obvious that the three main stakeholders blame each other, and there are additional problems in this world besides the dying fish. 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 along the river and analyze the water quality based on six indicators (e.g., 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 the activities and water-quality data are interrelated. The third mission focuses on reasoning about the data that's been collected, and drawing a preliminary scientific conclusion based on the hypothesis rendered in the preceding mission.

The fourth mission is set two 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 two 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 ulti-

mately 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 *not* been solved. 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 Park 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 again employ the time machine to travel five 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 Park in the future, students identify which changes occurred and how they reflect a socio-scientific solution.

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. Figure 4 shows the ECD models for a fragment of the ST competency (i.e., *Model the System*), with particular focus on "Create Causal Loop Diagrams."

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 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).

Competency Model

By the time students reach Mission 4 in Taiga Park, 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,

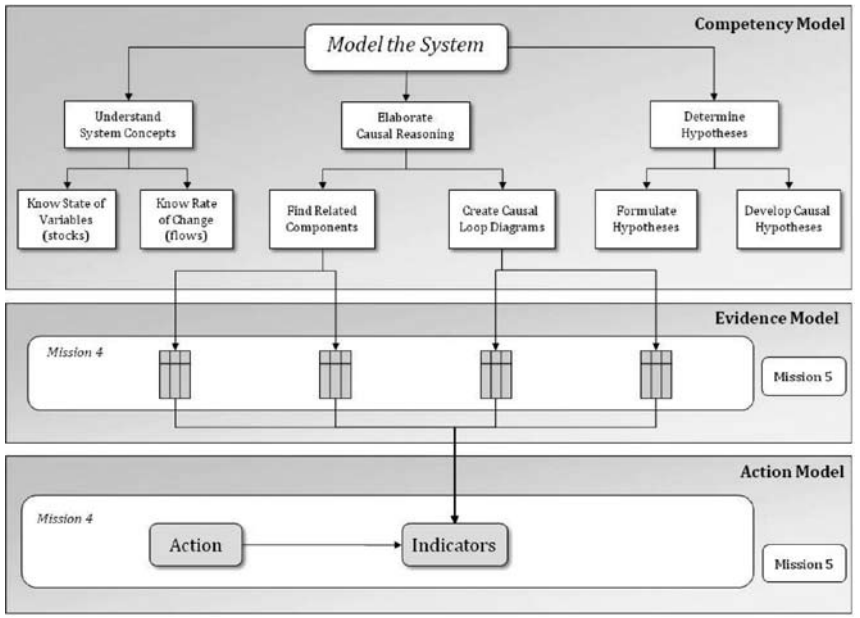

FIGURE 4
Conceptualization of ECD models applied to Taiga.

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 loop diagram that shows the system variables that are reducing the population of fish in the river.

Evidence Model

This model determines 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 quests such as diagrams created or short answers provided to specific questions, (b) rules for scoring the student submissions, and (c) weights in terms of the outcomes' contributions 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. Table 1 lists a few representative actions and their indicators relevant to various Taiga Park missions.

TABLE 1
List of a few relevant 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 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 • Correctly relate these activities to eutrophication
Create causal loop diagram	• Include complete set of variables and links in the diagram • Accurately identify relationships among variables (positive or negative)

In the current version of Taiga Park, 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. In addition to the essays, students can create and submit causal loop diagrams (demonstrating the relevant variables within the system and their cause-effect relationships). Within 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 to also 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. Additionally, there may be ambiguity in diagrams and subjectivity in assessing on the teachers' parts. We believe, however, that crafting causal diagrams is an important aspect of system thinking competency and this activity should be an integral (not optional) part of the game.

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 Park. In this illustration, we focus on an Excel-based software application called jMap (Jeong, 2008a; Shute, Jeong, & Zapata-Rivera, in press) that students can use to create causal diagrams. jMap is 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 or an aggregated group model (for more

information about the program, including links and papers, see: http://garnet.fsu.edu/~ajeong).

Using 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 to the current example). In jMap, comparisons between a student's and a target map begin by automatically coding/translating each map into a transitional frequency matrix (see Table 2). Each observed link within the student's map is recorded into the corresponding cell of the matrix.

Once all (i.e., student and expert maps) have been automatically coded into transitional frequency matrices, jMap regenerates visual diagrams/maps of each model using a standardized template to facilitate visual analysis and comparison of maps. Consequently, 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 (see Shute, Jeong, & Zapata-Rivera, in press for examples). jMap can also be used to aggregate all the frequencies across the

TABLE 2
Example of a transitional frequency matrix corresponding to Figure 5[3].

Transitional Frequency Matrix	Taiga Park income	Need more logging	Cutting trees	Soil erosion	Sediment in water	Temp. of water	Dissolved oxygen	Fish population
Taiga Park income		−1						
Need more logging			+1					
Cutting trees				+2				
Soil erosion					+3			
Sediment in water								−3
Temp. of water								
Dissolved oxygen								
Fish population	+2							

[3] In jMap, values of 1, 2, and 3 are used to denote differential relationship strengths, depicted in terms of line thicknesses. For simplicity, in our depiction (i.e., Figure 5) we kept all strengths the same. Values in Table 2 reflect more realistic strengths.

matrices of multiple learners to produce an aggregate frequency matrix representing the group. As a result, the resulting group map can also be superimposed over an individual learner's map or an expert map. Users (e.g., teachers, researchers, students, and others.) 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.

In this scenario, and as part of their gaming mission, students would draw their causal diagrams using jMap, which would contain a collection of system elements. 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.

Several validation studies have been conducted to test whether students can effectively create causal diagrams (e.g., Burns and Musa, 2001), and whether the diagrams make causal sense. For instance, Shute, Jeong, and Zapata-Rivera (in press) assessed students' abilities to construct causal diagrams. Students were asked to map out their theories/beliefs on how the choice of media (in an e-learning environment) shapes learning effectiveness. Students were given a set of 10 variables and asked to draw causal diagrams mapping out the relationships between the variables at three different points in time (spanning seven days). The study examined how causal maps evolved relative to understanding the issues. jMap was used to detect structural differences in students' causal diagrams as they progressed through the week. Jeong (2008) also showed that students are able to use causal diagrams to articulate and refine their beliefs on factors that impact collaborative learning and instructional strategies. In both studies, jMap demonstrated strong potential as a tool capable of (a) assessing students' understanding of complex phenomenon, (b) tracking evolving mental models, and (c) providing feedback to students' on their performance vis-a-vis an expert or reference diagram.

Adding stealth assessment to Taiga Park

Consider a student named Clara (which is a pseudonym for one of the authors of this paper). Two causal loop diagrams were obtained from her at two different points in time: during an early mission in Taiga Park, and then during her final mission. During the early mission, Clara blamed the decline-in-fish-population problem solely on the loggers. Her causal loop diagram at that point is shown in Figure 5. 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, and pH level of the water. The relationships between variables are also recorded directly in the diagram using a "+" (for a positive function) or a "−" (for an inverse function).

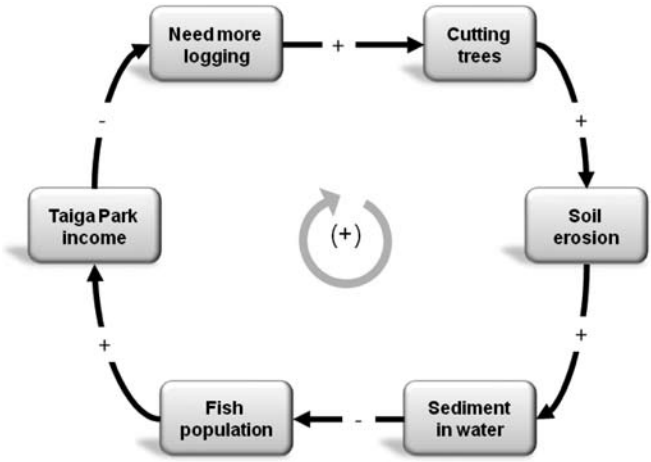

FIGURE 5
Clara's causal loop diagram at Time 1.

When she visits Taiga Park two years in the future, Clara quickly realized that her simple conceptualization of the problem (i.e., blaming just a single group of resident stakeholders – the loggers) and the ensuing solution (i.e., Ranger Bartle's banning the loggers from Taiga Park) was in vain. That is, two 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 Park (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 Figure 6) and recommends various regulations on all three stakeholders to Ranger Bartle.

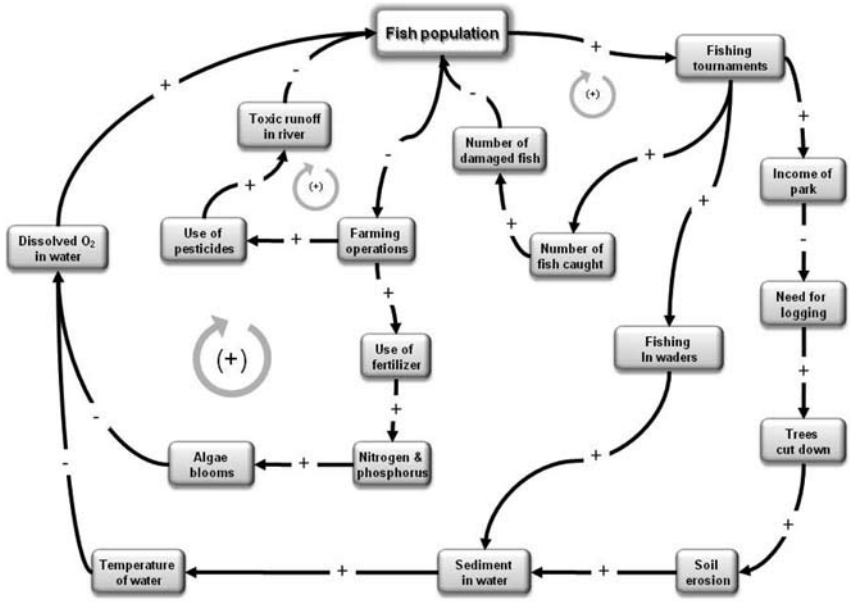

FIGURE 6
Clara's causal loop diagram—Time 2.

So how does jMap derive indicator values to feed into the Bayes net? Let's look at the jMap analysis comparing Clara's Time 1 map to an expert map. As shown in Figure 5, Clara demonstrated incomplete modeling of the system based on her performance on relevant indicators. A screen capture from jMap is shown in Figure 7. Here, jMap's automatically generated diagram uses colored links to visually identify differences between two selected maps – in this case between Clara's Time 1 map and an expert's 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. Black arrows represent positive relationships and grey ones represent negative relationships. By visual inspection, we can see that Clara has omitted three links and two important nodes in her causal loop diagram relative to the expert's map (shown by the three dashed arrows).

Clara's 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*

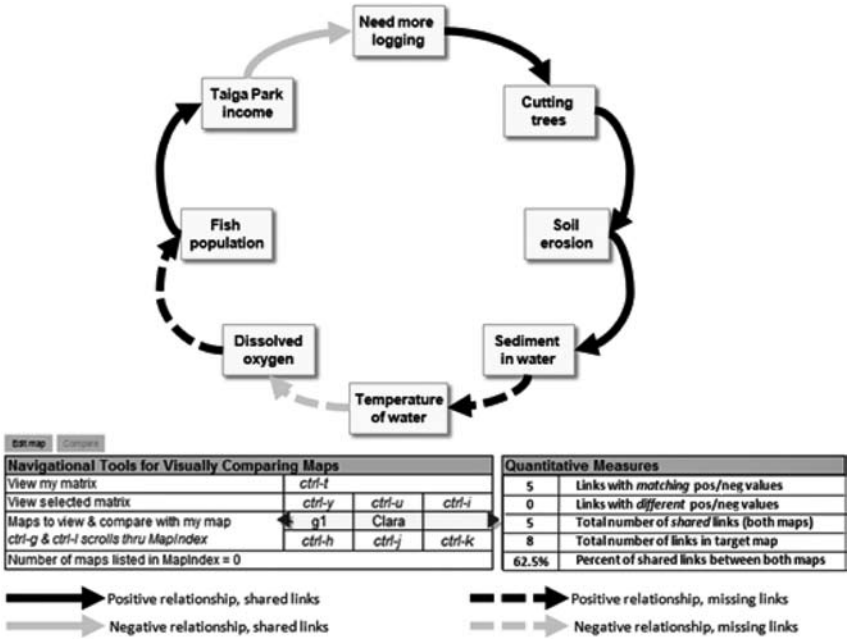

FIGURE 7
jMap interface showing a Clara's Time 1 map overlaid on the expert's map.

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"). In addition, the lab technician (or another knowledgeable character in Taiga) could provide feedback in the form of a causal loop diagram, like the one shown in Figure 7, explicitly including those variables in the picture. That way, she can see for herself what she'd left out.

In addition to the standardized maps, the jMap interface includes two tables, also shown in Figure 7. The table on the left includes navigational tools. 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, she would receive a "high" score on accuracy of links because she'd created the correct relations of the links in her diagram (i.e., positive vs. negative functions). These indicator outcomes are then inserted into the Bayes net (see Figure 8) through the Netica software that was used to create the Bayes net.

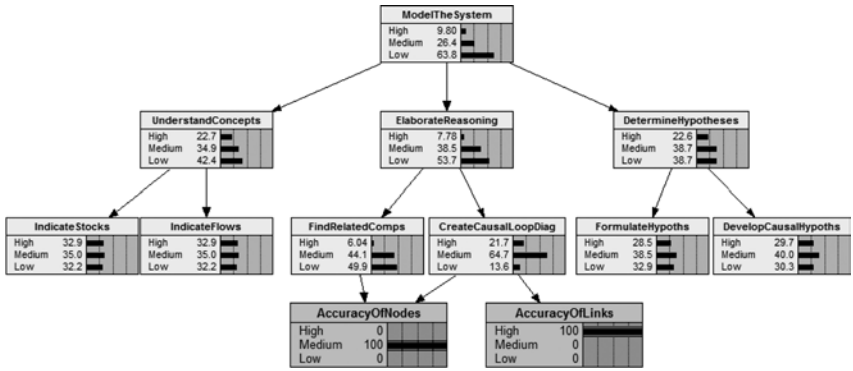

FIGURE 8
Bayesian model for Clara at Time 1.

Bayes networks provide a graphical demonstration to explain complex probabilistic relationships among different constructs (Pearl, 1988; Pearl & Russell, 2003). Bayesian analysis estimates directly and dynamically the value of quantity rather than the value of a test statistic (Reckhow, 2002). 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. For instance, Clara's Time 1 estimate for the competency, "create causal loop diagram" is medium; her "elaborate causal reasoning" competency, however, is estimated at low, as is her overall competency, "model the system." She has more work to do in Taiga Park, and this analysis and diagnosis targets particular areas for improvement.

By the final mission, as evidenced in her causal loop diagram shown in Figure 6, Clara has acquired a good understanding of the various systems operating in concert in Taiga Park. 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.

SUMMARY

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, and others.) being developed at the end of the ECD process. But in this paper, 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; (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 information gathering skill within a given 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 21st 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).

DISCUSSION

The main problem that we seek to address with this research is that educational systems (in the U.S. and around the world) need to identify ways to fully engage students through learning environments that meet their needs and inter-

ests (e.g., through well-designed educational games). We maintain that not only is it important to determine the skills needed to succeed in the 21st 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 challenge 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 non-human characters, can be used as the basis for assessment and support of valuable skills).

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 or provide valuable feedback to the learner.

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 (2010) 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.

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. Currently, Taiga Park 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 example, instead of spending countless hours grading essays and diagrams, teachers could simply review students'

competency models, and use that information as the basis to alter instruction or provide formative feedback (see Shute, 2008).

In conclusion, we propose using ECD, stealth assessment, and automated data collection and analysis tools to not only collect valid evidence of students' competency states in game environments, but to also reduce teachers' workload in relation to managing the students' work (or "play") products. 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 paper are intended to help teachers facilitate learning, in a fun and engaging manner, of educationally valuable skills not currently supported in school.

REFERENCES

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

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

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

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

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

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

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

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

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

Burns, J. R. & Musa, P. (2001). *Structural validation of causal loop diagrams.* Proceedings of the 19th International Conference if the Systems Dynamics Society, Atlanta, GA.

Calfee, R. C., & Valencia, R. R. (1991). *APA guide to preparing manuscripts for journal publication*. Washington, DC: American Psychological Association.

Choi, D., & Kim, J. (2004). Why people continue to play online games: In search of critical design factors to increase customer loyalty to online contents. *CyberPsychology & Behavior*, *7*(1), 11–24.

Chou, T.- J. & Ting, C.- C. (2003). The role of flow experience in cyber-game addiction. *CyberPsychology & Behavior, 6*(6), 663–675.

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

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.), *Lecture Notes in Computer Science*, (pp. 61–79). Berlin / Heidelberg: Springer.

Forrester, J. W. (1996). *System dynamics and K-12 teachers*. Retrieved August 08, 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. (2003). What *video games have to teach us about learning and literacy*. New York: Palgrave/Macmillan.

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

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

Gee, J. P. (2008). Video games, learning, and "content". In C. Miller (Ed.), *Games: Purpose and potential in education* (pp. 43–53). Boston, MA: Springer.

Jeong, A. C. (2008a). Discussion Analysis Tool (DAT). Retrieved December 22, 2008, from http://garnet.fsu.edu/~ajeong/DAT

Jeong, A. C. (2008b). *Assessing skills in scientific inquiry, argumentation, and causal modeling*. Paper presented at the 2008 American Educational Research Association conference for the Technology, Instructional, Cognition and Learning (TICL) Symposia, New York, NY.

Jensen, E. & Brehmer, B. (2003). Understanding and control of a simple dynamic system. *System Dynamics Review, 19*(2), 119–137.

Jonassen, D., Strobel, J., & Gottdenker, J. (2005). Model building for conceptual change. *Interactive Learning Environments, 13*(1/2), 15–37.

Lane, D. (2008). The emergence and use of diagramming in system dynamics: a critical account. *Systems Research & Behavioral Science*, *25*(1), 3–23.

Martin, L. A. (1997). Road Map 2: Beginner Modeling Exercise. MIT System Dynamics in Education Project. Retrieved August 05, 2008, from Massachusetts Institute of Technology (MIT), Systems Dynamics in Education Project Web site: 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*(2), 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*, *59*(4), 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 assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

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 Web site: 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.

Pearl, J., & Russell, S. (2003). Bayesian networks. In M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks* (pp. 157–160). Cambridge, Massachusetts: MIT Press.

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

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

Reckhow, K. H. (2002). Bayesian approaches in ecological analysis and modeling. In C. D. Canham, J. J. Cole, & W. K. Lauenroth (Eds.), *The Role of Models in Ecosystem Science* (pp. 168–183). Princeton, NJ: Princeton University Press.

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, N.H.: High Performance Systems, Inc.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment*, *8*(4). Retrieved from http://escholarship.bc.edu/jtla/vol8/4

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

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

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

Senge, P., Cambron-McCabe, N., Lucas, T., Smith, B., Dutton, J. & Kleiner, A. (2000). *Schools that learn: A fifth discipline fieldbook for educators, parents, and everyone who cares about education*. New York: Doubleday/Currency.

Simon, H. A. (1996). *The sciences of the artificial.* Cambridge, Mass: 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). New York, NY: 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. (in press). Stealth assessment in computer-based games to support learning. To appear in S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction.* Charlotte, NC: Information Age Publishers.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it - Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in 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, NY: 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., Rieber, L., & Van Eck, R. (in press). Games . . . and . . . learning. To appear in R. Reiser & J. Dempsey (Eds.), *Trends and issues in instructional design and technology, 3rd Edition*, Upper Saddle River, NJ: Pearson Education, Inc.

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.

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. H. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24*(3)*,* 223–258.

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

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

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press. Published originally in Russian in 1930.

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

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.