Assessment of Student Learning in Science Simulations and Games

Edys S. Quellmalz, Michael J. Timms, & Steven A. Schneider
WestEd

Introduction

The development of science simulations and games has outpaced their grounding in theory and research on learning and assessment. Both genres draw upon the affordances of technology to provide students with opportunities to engage in active exploration and experimentation in virtual science environments that are impractical or impossible to access otherwise. Science simulations typically fall into three categories—simulations of science phenomena, multiuser virtual environments, and virtual laboratories. Educational, or "serious," games immerse learners in realistic science worlds as the students gain and use knowledge to solve mysteries or to advance through multiple levels. Learning objectives may be acquisition of declarative knowledge—scientific facts, concepts and principles; formation of schema—connected knowledge structures about science systems; acquisition of procedural knowledge—steps in using tools and equipment; or strategic knowledge—learning when to employ inquiry skills and hone model-based reasoning. Even broader goals may target metacognitive strategies and epistemic strategies for far transfer.

Simulations and games typically present tasks that are generally interactive, requiring the student to construct understandings and conduct iterative investigations within the virtual environments. Activities may vary from set procedures to graduated levels of complex strategies. Tasks may provide feedback and hints to scaffold learning progress.

Relatively scarce, however, are clearly articulated descriptions of the evidence gathered to support claims of student learning. In most instances, rich streams of data from interactive tasks are not tapped as evidence of learning. Assessments of learning from simulations and games often resort to paper-based conventional task and item formats with limited possibilities for measuring the significant kinds of complex science learning targeted.

Goals

The goals of the paper are to:

- Explore the promise and potential of games and simulations for science learning and as environments for formative, classroom assessments and for summative, large scale assessments;
- Summarize research and practice on learning assessments in a range of educational science simulations and games;
- Use the conceptual assessment framework of evidence-centered design to analyze current assessments in science simulations and games; and
- Propose an agenda for research on assessments of science learning in and with simulations and games.

Approach

This paper will focus the lenses of learning and assessment theory and research on the current state of practice employed in assessing learning in science simulations and games. The paper will compare the range of complex knowledge and processes advocated in challenging national and international science standards with the types of learning outcomes addressed in science simulations and games. The conceptual assessment framework of evidence-centered design (ECD) will be employed to analyze the types of student models targeted by K-16 science simulations and games, the features of task models used to elicit evidence of learning, and the evidence models employed, including psychometric methods, to analyze and report student learning.

Cognitively-Principled Assessment Design

In the domain of science, core knowledge structures are represented in models of the world built by scientists (Hestenes et al., 1992; Stewart & Golubitsky, 1992). Technologies are seen as tools that support schema formation by automating and augmenting performance on cognitively complex tasks (Norman, 1993). The NRC report, *Knowing What Students Know*, presents advances in measurement science that support the integration of cognitive research findings into systematic test design frameworks. Evidence-centered assessment design involves relating the learning to be assessed, as specified in a *student model*, to a *task model* that specifies features of the task and questions that would elicit observations of learning, to an *evidence model* that specifies the student responses and scores serving as evidence of proficiency (Messick, 1994; Mislevy et al., 2003; Pellegrino et al., 2001). These three components of the conceptual assessment framework provide a structure for evaluating the state of current assessment practices in science simulations and games.

Below, we address assessments of science learning in the three types of simulations, then in games. We discuss the psychometric issues related to assessments in these innovative environments, followed by suggested research strategies for furthering the quality and utility of science learning assessments in simulations and games.

Assessments of Science Learning in Simulations

In this paper, we define science simulations as dynamic representations of spatial, temporal, and causal phenomena in science systems that learners can explore and manipulate. In contrast to animations, where students view predetermined scenes and can only control viewing direction and pace, simulations adapt the dynamic displays in response to learner inputs. Key features of simulations include manipulation of structures and patterns that otherwise might not be visible or even conceivable, representations of time, scale, and causality, and the potential for generating and superimposing multiple physical and symbolic representations. Moreover, simulations have the potential to illustrate content in multiple representational forms, which can strengthen students' mental models of concepts and principles and also reduce potentially confounding language demands. Simulations can present the opportunity for students to engage in the kinds of investigations that are familiar components of hands-on curricula and also to explore problems iteratively and discover solutions that students might not have discovered in other modalities. Importantly, simulations also can make available realistic problem scenarios that are difficult or impossible to create in a typical classroom.

The most prevalent forms of simulations are two-dimensional computer simulations of science phenomena and virtual laboratories that simulate on-screen the experiments that are traditionally performed in real school laboratories. Virtual laboratories are valued for savings on equipment costs, ease of logistics, and safety. Virtual labs can be very efficient also by allowing several repetitions of an experiment in limited time. The technology platform offers data collection advantages, with students able to capture, record and analyze data easily and with time efficiency.

Another form of simulation is the three dimensional, multi-user virtual environment (MUVE) that constructs simulated immersive experiences. In a MUVE, each user has a virtual representation, called an avatar, and moves this graphical avatar through a three dimensional, virtual world. In addition to the benefits attributed to simulations, such as situating learning in a more authentic context and providing direct experiences and interaction with intangible, abstract, ideal, complex, or otherwise unavailable scientific phenomena, multiuser virtual environments permit learners to customize the learning environment and engage in collaborative problem-solving.

Research on Simulations and Student Science Learning

Numerous studies illustrate the benefits of science simulations for student learning. The benefits of simulations have been studied by multimedia researchers, by cognitive psychologists, by curriculum developers, and by commercial companies.

Learning from Simulations. Multimedia research suggests that when degrees of learner control and interactivity are variables, spatial representations seem to enable effective mental models and visualizations (Schwartz & Heiser, 2006). Rieber et al. (2004) found that students given graphical feedback during a simulation on laws of motion with short explanations far outperformed those given only textual information. Simulations have been shown to facilitate knowledge integration and a deeper understanding of complex topics, such as genetics, environmental science, and physics (Horwitz et al., 2007; Hickey et al., 2003; Krajcik et al., 2000; Doerr, 1996). Model-It was used in a large number of classrooms, and positive learning outcomes based on pretest-posttest data were reported (Krajcik et al., 2000). Ninth-grade students who used Model-It to build a model of an ecosystem learned to create “good quality models” and effectively test their models (Jackson et al., 1996). After participating in the Connected Chemistry project, which used NetLogo to teach the concept of chemical equilibrium, students tended to rely more on conceptual approaches than on algorithmic approaches or rote facts during problem solving (Stieff & Wilensky, 2003). Seventh, eighth-, and ninth-grade students who completed the ThinkerTools curriculum performed better than high school students on basic physics problems, on average, and were able to apply their conceptual models for force and motion to solve realistic problems (White & Frederiksen, 1998). An implementation study of the use of *BioLogica* by students in eight high schools, showed an increase in genetics content knowledge in specific areas, as well as an increase in genetics problem-solving skills (Buckley et al., 2004). Research conducted with the Modeling Across the Curriculum project measured inquiry skills *in situ.* Log files of student responses correlated systematic (vs. haphazard) inquiry performances with overall learning gains (Buckley et al, 2009). At the middle school level, a simulation of an aquatic ecosystem was used to allow students to look beyond the surface structures and functions they could see when an aquarium served as a physical model. The simulation allowed students to create connections between the macro-level fish reproduction and the micro-level nutrification processes (Hmelo-Silver, et al., 2008).

Commercial simulation packages are becoming more prevalent in schools. Seventy-seven simulation products for middle and high school are currently being reviewed in an NSF-funded synthesis project (Scalise, et al., 2009). The simulations span topics such as thermodynamics, chemistry, genetics, and cell structure and function. The most common evaluation method, used in slightly more than half of the products reviewed, was a pre-post comparison of student learning on goals and objectives. Approximately four percent of the studies reported no gain, about 25 percent reported mixed outcomes in which some groups showed learning gains but others did not, just over 20 percent reported gains under the right conditions, and about 51 percent reported overall gains.

At the post secondary level, the Physics Education Technology (PhET) project conducted over 275 individual student interviews during which the college undergraduates described what they were thinking as they interacted with over 75 simulations for use in teaching undergraduate physics, chemistry, and physical science (Adams, et al., 2008). The researchers observed that the simulations were highly engaging and educationally effective.

Virtual Labs. Virtual laboratory products, such as Model Chemlab, Biology Labs Online, and Virtual ChemLab, are becoming increasingly used in high schools, particularly smaller schools in rural areas that do not have science labs. At the middle school level, virtual labs have been used to investigate topics such as density, porosity/permeability, and plant growth. The labs allow students to use virtual microscopes or examine different types of rocks. At the high school level, there are many virtual lab products designed for use in chemistry, physics, and biology courses. Students can perform virtual experiments dealing with aqueous chemistry, such as acid-base reactions and solubility; physics experiments involving force and motion, springs, and electrical circuits; and virtual dissections of frogs and other animals. Virtual labs are also being used in introductory level college science and engineering courses to prepare students for work in real world labs in fields such as thermodynamics, robotics, and biotechnology.

A review of literature comparing hands-on, virtual and remote laboratories in university science education found mixed results and that researchers were confounding many different factors, perhaps over-attributing learning success to the technologies used (Ma and Nickerson, 2006). The review in progress of 25 commercially available virtual labs also found mixed evidence of effectiveness, but that the majority of the products produced learning gains (Scalise, et al., 2009).

Multi-User Environments. Studies of science learning in multi-user virtual environments are fewer, but promising. A project using "collective simulations" allowed students to learn about the intricacies of interdependent complex systemsby engaging in discourse with other students and teachers (Repenning & Ioannidou, 2005). The infrastructure created immersive learning experiences based on wirelessly connected handhelds. As part of the Mr. Vetro human body systems simulation prototype, each group controlled physiological variables of a single organ on their handheld computer. A central simulation gathered and projected all the data. Students subjected Mr. Vetro to different levels of exercise and controlled the heart and lungs to optimize his physical condition. Students recorded their data and used their parameter values to reach conclusions and answer paper-based questions prepared by their teacher. The project found that while students had some understanding of each separate organ, they did not have a clear sense of the connection between the circulatory and the respiratory systems. The pilot study focused, however, on user testing of the technologies and did not report student learning following the collaborative activities.

Another study of multi-user virtual environments examined students' understanding of a virtual infectious disease in relation to their understanding of natural infectious diseases. Two sixth-grade classrooms of students between the ages of 10 and 12 (46 students) took part in a participatory simulation and completed pre and post surveys about virtual infectious diseases (Nulight, et al, 2007).

Analysis of Assessments of Science Learning Promoted*By* Simulations

In very few instances were assessments of student learning actually embedded within the simulation or designed to take advantage of its technological capabilities. In most of the projects using science simulations, paper-pencil tests were used and details of their design, i.e., science knowledge and inquiry assessed, types of tasks and items, were not described in study publications, technical quality. This was the case with research-based simulation projects, curriculum development projects, and with the commercial simulations and virtual labs.

Criteria for Establishing the Quality of the Assessments. For K-12 simulations and assessments, the *first* criterion for the quality of the assessments (and simulations) would be documentation of the alignment of the targeted science content knowledge and inquiry skills with consensus national science standards such as the 2009 NAEP Science Framework, the National Science Education Standards, and the AAAS Benchmarks for Scientific Literacy. For simulations at the post secondary level, alignment should be documented of clearly specified learning goals and the assessments intended to measure them. Of particular importance would be the extent to which the assessment items provided evidence that the simulations promoted deep science understanding and inquiry. Since simulations typically address only a limited number of content and inquiry goals, are the simulation targets teaching and testing standards less well addressed by paper formats? And, do the assessment items and tasks, in turn, address the full range of a simulation's goals? A *second* criterion would be documentation of the quality of the assessment items and tasks. Within the evidence-centered design framework, are the knowledge and skills in the student model elicited by items and tasks representing the task model, and do the data from the assessment tasks and items actually provide evidence of the simulation goals? An inherent dilemma is the disconnect between the science knowledge and skills that can be tested in static, paper item formats in comparison to the science knowledge and skills that can be tested within the dynamic simulation environment. Are technical qualities of the assessment's item characteristics, reliability and validity reported? A *third* criterion is the utility of the assessments for the assessment purpose. If the assessments were intended for summative purposes, was the intent to document student learning from just the simulation? If the assessments are intended to inform instruction, did the study determine if instructors found the assessment information useful and did they act on it? In many of the reports, the assessments were used to document the effectiveness of the simulation, but were developed and used for research and evaluation purposes and not apparently intended to be included as a component of the simulation learning activity. The ultimate validity of the assessments must rest on the adequacy of the assessments for supporting the intended inferences and actions.

Potential of Simulation-based Assessments. Assessment tasks *within* science simulations can elicit evidence of rich, principled science learning. Although simulations can present assessment tasks and questions that ask for the same basic foundational knowledge often tested by paper-based items—such as definitions of consumers, producers, and producers in an ecosystem—more importantly, simulations can test students' knowledge of how components of a system interact (e.g. flow of energy in food webs) and also assess understandings of emergent model behaviors (e.g., predator-prey effects on population dynamics). Moreover, simulations can directly assess students' abilities to *conduct* inquiry in tasks such as requiring observations of organisms in a novel ecosystem to determine their roles and interrelationships. Simulations also permit iterative investigations of the impacts on population dynamics of multiple variables changing at the same time (Quellmalz, Timms, & Buckley, in press). Because simulations use multiple modalities to represent science systems and to elicit student responses, students with diverse learning styles and language backgrounds may have better opportunities to demonstrate their knowledge than are possible in text-laden print tests (Kopriva, et al., 2009).

Status of Assessments of Learning from Science Simulations. Although the use of simulations in the projects reviewed generally was positively related to student science learning outcomes, the assessments typically did not take advantage of the capabilities of the simulation technology to gather evidence of active inquiry skills and model-based reasoning. Most science simulation projects did not assess the full range of knowledge and inquiry skills supported by the simulation environment. For instance, the pretest and posttest for Model-It were paper pencil-based with open-ended and multiple-choice items. Similarly, assessment data collected in the 2004 BioLogica study included pre- and posttest paper-pencil data as well as log files showing types of student usage. Similarly, in the multi-user simulations, paper based items were used to test science learning.

Learning effectiveness of the PHeT simulations was examined during think-alouds and interviews of users (Adams, et al., 2008). The students were asked prediction-type conceptual questions, then, during or after interacting with the simulation, they were allowed to revise their answer. However, the report of the interviews did not specify the science knowledge required, any inquiry skills involved, nor the interview questions asked about learning. The report indicated that most students understood the concepts covered in the simulation well enough to explain them accurately and to use them to make accurate predictions, a level of understanding far beyond what the researchers had observed was typically obtained from the coverage of these concepts in a physics course. Much of the report focused on design features of the simulations related to their usage. Although the cognitive interviews were better methods for tracing student understanding and reasoning, the simulations were not designed to incorporate assessments as components of course implementations.

In a few projects, researchers collected evidence of student learning in items and tasks embedded in the simulations. The types of tasks and items included not only conventional selected and constructed written responses, but other forms of constructed responses as well, such as drawing arrows to depict the direction and magnitude of a force or the flow of energy in a food web. More innovative response formats included analyses of sequences of problem solving actions. For example, the Modeling Across the Curriculum project used multiple-choice pre and post tests aligned to the learning goals of the project to measure learning gains. In addition, however, increasingly complex problem-solving or inquiry tasks with fading scaffolding were included at the end of each learning module. Performance on these tasks was measured by analyzing the sequence of students' problem solving actions in addition to their explicit answers. Task performances were significantly correlated with overall learning gains to various degrees (Buckley, et al., 2009).

In sum, reports of the educational effectiveness of science simulations tend not to describe in sufficient detail the science knowledge and inquiry targeted in the designs of assessment items and tasks administered, and the analyses used to draw conclusions about learning. The power of simulation environments is that, once developed, they can host a large number of assessment questions and inquiry tasks ranging from tests of component concepts to integrated knowledge in extended investigations. The affordances of the simulations should be harnessed to design assessments that capture the value added of the levels of complexity of knowledge, reasoning, and inquiry the simulations address.

Use of Science Simulations *For* Assessment

An emerging function of simulations is their use as innovative assessment formats. Designs of simulations for summative purposes can aim to elicit evidence of inquiry skills and use of science principles by setting forth sets of tasks linked to scenarios about science phenomena or laboratory investigations. Use of science simulations for formative purposes can take advantage of the technology's capabilities for providing individualized feedback and hints to scaffold and benefit learning.

Summative Assessment Uses of Science Simulations. The most prevalent use of technology in large-scale testing currently involves support for assessment logistics related to online delivery and automated scoring. Accountability programs remain caught up in perceived needs to document the comparability of scores and interpretation of results with computer-based and paper forms. A new generation of assessments, however, is attempting to break the mold of traditional testing practices. Innovative assessment formats, including simulations, are being designed to measure complex knowledge and inquiry previously impossible to test in paper-based or hands-on formats. The new generation of testing is reconceptualizing assessment design and use and aiming to align summative assessment more directly to the processes and contexts of learning and instruction (Quellmalz & Pellegrino, 2009).

A number of large-scale testing programs have begun to design innovative problem sets and item types that promise to transform traditional testing. The area of science assessment is pioneering the exploration of innovative problem types and assessment approaches across K-12. In 2006, the Programme for International Student Assessment (PISA) pilot tested the Computer-based Assessment of Science (CBAS) with the aim of testing science knowledge and inquiry processes not assessed in the PISA paper-based test booklets. CBAS tasks included scenario-based item and task sets such as investigations of the temperature and pressure settings for a simulated nuclear reactor. The 2009 National Assessment of Educational Progress (NAEP) Science Framework and Specifications proposed designs for Interactive Computer Tasks (ICT) to test students' ability to engage in science inquiry practices. These innovative formats were included in the 2009 NAEP science administration. At the state level, Minnesota has an online science test with tasks engaging students in simulated laboratory experiments or investigations of phenomena such as weather or the solar system. Large-scale testing programs are beginning to explore the possibilities of dynamic, interactive tasks for obtaining evidence of science learning achievement levels. The current accountability stakes and constraints, however, tend to restrict a program's options to take full advantage of the computer's ability to provide tasks adapted to an examinee's performance during the test. Nonetheless, science simulations open significant opportunities for the design of assessments of systems thinking, model based reasoning, and scientific inquiry advocated in national science standards, but seldom tapped in paper-based tests (Quellmalz et al., 2005).

Formative Assessment Uses of Science Simulations. Simulations are well-suited to supporting some of the data collection, complex analysis, and individualized feedback and scaffolding features needed for formative assessment (Brown, Hinze & Pellegrino, 2008). The computer's ability to capture student inputs permits collecting evidence of processes such as equipment use, problem solving, and strategy use as reflected by tool features manipulated, information selected, sequence of trials, numbers of attempts, and time allocation. Many of the design and practical limitations of systematic uses of formative assessments in classrooms can be overcome by the use of technology to align, design, deliver, adapt, and score assessments within rich task environments that measure deep understandings in a feasible and cost-effective manner (Quellmalz & Haertel, 2004).

In an ongoing program of research and development, WestEd's SimScientists projects are studying the suitability of simulations as environments for formative and summative assessment and as curriculum modules to supplement science instruction. (Quellmalz, et. al, 2008; Quellmalz, Timms & Buckley, in press). One of the SimScientists projects, Calipers II, funded by the National Science Foundation, is studying the use of science simulations for end-of-unit, summative, benchmark purposes and for curriculum embedded formative purposes. Students use the simulation to investigate science problems that relate to understanding increasingly complex levels of grade-appropriate models of science systems. A learning progression underlies a sequence of assessments of components and roles of organisms in a system, to interactions among components in a system, to emergent behaviors of interactions in a system (Buckley, et al., submitted). To assess transfer, students conduct a range of inquiry activities across different ecosystems.

Figure 1. *Screenshot of SimScientists Ecosystems Benchmark Assessment Showing a Food Web Diagram Interactively Produced by a Student After Observing the Behaviors of Organisms in the Simulated Australian Grasslands Environment.*

Figures 1 & 2 present screen shots of tasks in a SimScientists summative, benchmark assessment designed to provide evidence of middle school students' understanding of ecosystems and inquiry practices after completion of a regular curriculum unit on ecosystems. Students are presented with the overarching problem of preparing a report to describe the ecology of an Australian grasslands for an interpretive center. They investigate the roles and relationships of the animals, birds, insects, and grass by observing animations of the interactions of the organisms. Students draw a food web representing interactions among the organisms in the novel ecosystem. The assessment then presents sets of simulation-based tasks and items that focus on students' understanding of the emergent behaviors of the dynamic ecosystem by conducting investigations with the simulation to predict, observe, and explain what happens to population levels when numbers of particular organisms are varied. In a culminating task, students present their findings about the grasslands ecosystem.

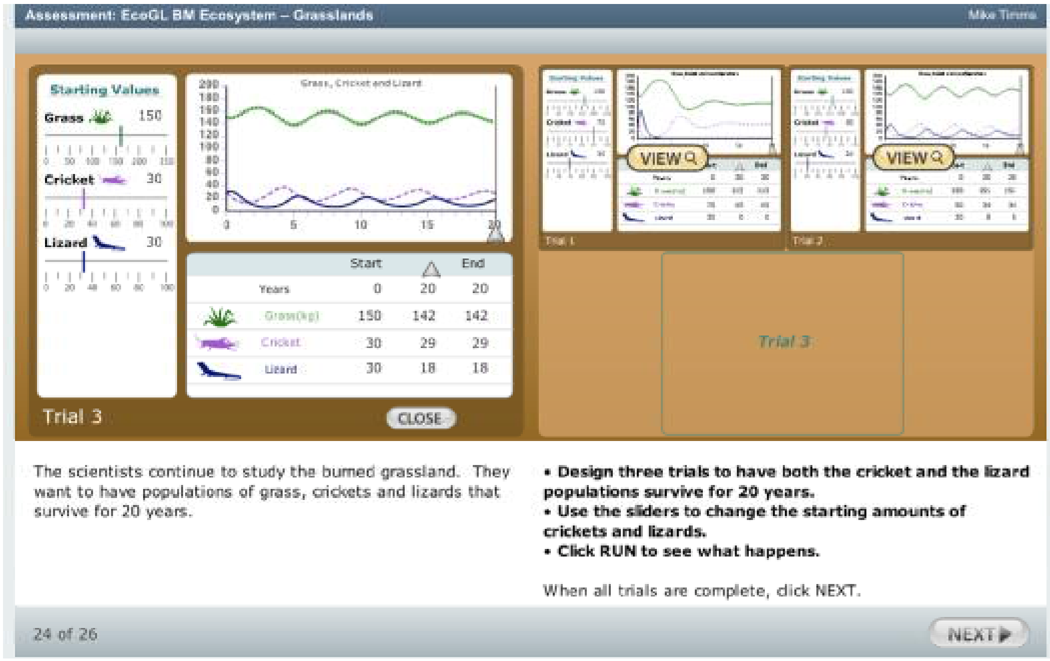

Figure 2. *Screenshot of SimScientists Ecosystems Benchmark Assessment Showing a Student's Investigations with the Interactive Population Model*

In a companion set of curriculum embedded, formative assessments situated in a different ecosystem, a mountain lake, the technological infrastructure identifies types of errors and follows up with feedback and graduated coaching. Levels of feedback and coaching progress from identifying that an error has occurred and asking the student to try again, to showing results of investigations that met the specifications. Figure 3 shows a screenshot in which feedback has been provided as the student is constructing a food web diagram after they have observed the organisms interacting in the simulated mountain lake environment. Students self-assess their constructed responses by judging if their explanations meet criteria or match a sample response.

These examples illustrate ways that assessment tasks can take advantage of the affordances of simulations to present significant, challenging inquiry tasks, provide individualized feedback, customize scaffolding, and promote self-assessment, metacognitive skills. Reports generated by the system for teachers and students indicate the level of additional help students may need and classify students into groups for tailored follow on, off line reflection activities.

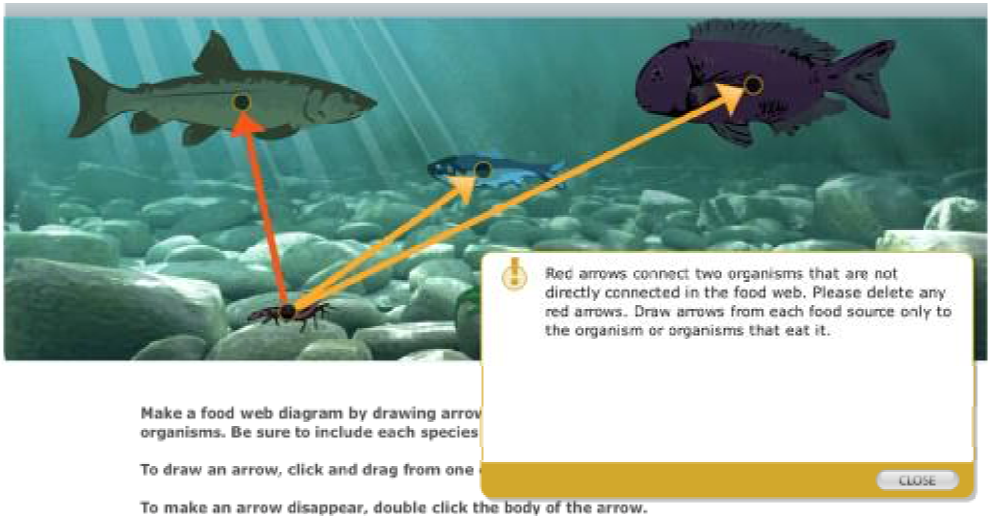

Figure 3. *Screenshot from the SimScientists Ecosystems Mountain Lake Embedded Assessment Showing Dynamic Feedback to a Student on the Food Web Diagram That She is Constructing*

Summary and Recommendations for Simulation-Based Assessments

Science simulations hold enormous promise for measuring significant science learning. Not only can simulations represent temporal, causal, dynamic and unobservable phenomena, the technology can permit a greater range of ways that students can express their understandings and inquiry strategies. Efficient data capture can allow real time (on-the-fly) aggregation of responses, individualized feedback, customized coaching, and adaptive tasks. However, the project reports reviewed for this paper did not provide detailed descriptions of the student, task and evidence models of assessments used to document student learning from science simulations. Analyses were not possible of the depth of understanding of the science content or the nature of scientific inquiry tested. Descriptions of the source of the items or data on the item and task quality were not typically provided, therefore the evaluations of the technical quality of the assessments are next to impossible. Examples of simulation-based formative assessments were not found in the published literature. Most technology-based formative assessments tend to focus on declarative knowledge, conventional multiple-choice formats, and offer little follow-on scaffolding. Nor, do technology-based assessments take advantage of cueing techniques such as highlighting and movement to offer worked examples of successfully completed tasks.

If science simulations are to become components of science curricula and widely used, documentation of the dependent measures of the simulations' effects on learning will be essential, along with data on the technical quality of the dependent measures. Moreover, if science simulations are to become environments for testing complex science learning, rigorously designed specifications of student models, task models, and evidence models will need to be developed and documented. Whether science simulations are used for formative or summative assessment purposes, evidence of the technical quality of tasks, items, and their aggregations must be reported, as would be required for the dependent measures in scientific research. The transformation from traditional to innovative assessment will need to be supported by credible evidence of its quality.

Assessment in Science Educational Games

Using the Affordances of Games for Science Education. Gee (Gee, 2007) likens the strategic thinking and problem solving required in popular commercial games to the best sorts of science instruction that occur in schools today. It is these qualities that have led to the growth of the field of educational or "serious" games that is attempting to apply principles from the consumer gaming industry to the development of engaging games that help children learn (Beal et al, 2002). In consumer games, the primary purpose is entertainment, but serious games as discussed here, have the purpose of education or training. There are characteristics of serious games that make them especially effective in science learning, and in 2005, the NSF sponsored The National Summit on Educational Games with The Federation of American Scientists and the Entertainment Software Association. The summit found that games offer several attributes important for learning—clear goals, tasks that can be practiced repeatedly until mastered, monitoring learner progress and adjusting instruction to learner level of mastery, closing the gap between what is learned and its use, motivation that encourages time on task, and personalization of learning. In addition, the types of skills that players are required to master include processes and skills such as strategic and analytical thinking, problem solving, planning and execution, decision-making, and adaptation to rapid change. These are skills that are relevant to science education. Games also provide the chance for students to apply practical skills they have learned, work in teams, and to train for high-performance situations in a low-consequence-for-failure environment.

Examples of Science Games and Assessments. Although games seem to offer an opportunity to enhance students' learning of complex science principles, research on how to effectively assess student learning in or with game environments is still in its infancy. A number of serious science games have been developed that provide rich environments. Three recent examples of such games include Quest Atlantis: Taiga Park (Barab, 2006; Barab & Dede, 2007a; Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007b; Hickey, Ingram-Goble, & Jameson, 2009), River City (Ketelhut, Dede, Clarke, Nelson, & Bowman, 2008), and CRYSTAL ISLAND (Rowe, McQuiggan, Robison, & Lester, 2009). While these games provide students with simulated environments in which to apply science inquiry skills, the assessments of learning elements in them are not as developed as the game play elements. Work is underway, however, to conceive of and apply more dynamic assessment methods suited to the highly interactive game environments, as the following discussions of each of the three examples illustrate. The examples are not meant to be exhaustive of the field of serious science games, and all three examples are of a similar type, but they are useful for examining the current state of the art in assessment and serious games.

The first example, Quest Atlantis: Taiga Park, is part of Quest Atlantic, an international learning and teaching project that uses a 3D multi-user environment to immerse children, ages 9-16, in educational tasks. It aims to combine strategies used in commercial games with principles from educational research on learning and motivation. In the Taiga Ecological Sciences Curriculum, students perform quests in the Taiga Park environment, which includes a river, loggers, tourists, an indigenous farming community, a fishing resort, and park administration. This simulated world was designed to engage students in complex socioscientific inquiry while also helping them learn ecological science concepts like erosion, eutrophication, and hypothesis testing. Students' avatars interact with virtual characters and data in order to evaluate competing explanations for declining fish populations in the Taiga River. However, the assessment of students is undertaken by classroom teachers who score the written mission reports submitted by students (Hickey et al., 2009). To date, the assessment is not embedded in the game play, although Shute et al. (Shute et al., 2009) use Quest Atlantis: Tiaga Park as the basis for a theoretical example of how evidence-centered design (ECD) could be used to develop assessments that would include “stealth” assessments and Bayesian Networks to monitor student progress and provide more automated feedback to students and teachers. Stealth assessment is Shute’s term for unobtrusive assessments that are so seamlessly embedded in game play that they are not noticed by the student playing the game. Shute also proposes that in games, the ECD task model can be conceived as an “action model,” reflecting the fact that the object of assessment in the game is to dynamically model the students’ actions rather than modeling the tasks they did which does not imply the same dynamic measurement.

Similarly, River City is a game-like multiuser virtual environment in which students can represent themselves through avatars, access various locations, use digital tools like microscopes, and undertake collaborative tasks in order to solve a mystery (Ketelhut et al., 2008). The River City virtual world is set in the 1800s and has a river, catchment area, and a town with institutions like a hospital and a university. Students interact with one another as avatars, or with computer-based agents, digital objects, and instructor avatars. Students work in teams to develop hypotheses regarding one of three different illnesses that have infected the town. At the end of the investigation, there is self-assessment as students compare their research findings with those of other teams in their class in order to discuss potential hypotheses and causal models for the diseases. There is also some rule-based assessment that is implemented in an embedded individualized guidance system that uses personalized interaction histories collected on each student’s actions to offer real-time, customized support for science inquiry. Ketelhut et al. found that the use of the individualized guidance system had a statistically significant positive impact on students gain scores on a test of content knowledge in science inquiry and disease transmission. This effect was more positive for female students than males. The researchers plan to collect more detailed data on when and if the students first use the guidance system, which messages they view, where they are in the game when they view them, and what they do next. They hope that this will allow them to provide more information to teachers on how students are progressing. Although some automation of assessment has been made in River City, the primary scoring of content knowledge and skills has, to date, relied upon human scoring using rubrics after the students have finished working in the environment.

CRYSTAL ISLAND is a narrative-centered learning environment built on Valve Software’s Source™ engine, the 3D game platform that is used for the commercial game Half-Life 2. On CRYSTAL ISLAND, students play the role of Alyx, the protagonist who is trying to discover the identity and source of an unidentified infectious disease. Students move their avatar around the island, manipulating objects, taking notes, viewing posters, operating lab equipment, and talking with non-player characters to gather clues about the disease’s source. Settings on the island include a beach area with docks, a field laboratory, underground caves, and a research camp. To progress through the mystery, students must explore the world and interact with other characters while forming questions, generating hypotheses, collecting data, and testing hypotheses. As students work to solve the mystery, they work through five different aspects of the science curriculum related to diseases. In the first two problems students deal with pathogens, including viruses, bacteria, fungi, and parasites. Students interact with in-game experts, books and posters to gather information that will enable them to trace the cause of the recent sickness among the scientists on the island. In the third problem, students have to compare and contrast their knowledge of four types of pathogens. In the fourth problem, students work through an inquiry-based hypothesis-test-and-retest problem in which they complete a “fact sheet” on the disease and have it verified by the camp nurse. In the final problem, the student selects an appropriate treatment plan for the sickened CRYSTAL ISLAND researchers.

Assessment in CRYSTAL ISLAND is evolving. Currently the assessment is mainly embedded in the reaction of in-game characters to the student's avatar. Researchers have been gradually building pedagogical agents in the game that attempt to gauge the student's emotional state while learning (anger, anxiety, boredom, confusion, delight, excitement, flow, frustration, sadness, and fear) and react with appropriate empathy to support the student's problem solving activities (McQuiggan, Robison, & Lester, 2008; Robison, McQuiggan, & Lester, 2009). A note taking feature was added and student notes were scored by researchers using rubrics (McQuiggan, Goth, Ha, Rowe, & Lester, 2008). Student notes were coded into four categories. One category was for student notes that contained facts from the narrative storyline, such as comments on the plot, objects, or symptoms of illness of particular characters. The second category was for facts from the curriculum, such as definitions or characteristics of viruses and bacteria. The third category was for student notes that explicitly expressed possible solutions regarding the source or cause of the outbreak or solution to the scientific mystery. A fourth category was used for student notes that proposed a hypothesis that was either narrative (e.g., suspecting a character of poisoning others) or curricular (e.g., guessing the cause of the disease wreaking havoc on the island). A fifth category of student notes dealt with lists of tasks they wanted to complete. McQuiggan et al. found that students who took hypothesis notes performed better on the posttests of content knowledge, which seems to indicate that it is important to scaffold students' hypothesis generation activities in such a learning environment. There were, however differences among the students, with girls taking more notes than boys, and high-mastery students taking more notes too. McQuiggan et al. also investigated if machine learning techniques could be applied to create measurement models that successfully predict the note-taking categories characterizing the content of the notes produced by human scorers. They found that the best performing model was support vector machines (SVM), a set of related supervised learning methods used for classification, which correctly classified 86.4% of instances. The SVM model was followed by naïve Bayes Net (83.2%), nearest neighbor (81.0%), a pattern analysis method for classifying objects based on how close they are to training examples, and decision tree (80.6%), which indicates that a variety of methods could be applied to score such notes in real time.

***Analysis of Assessments of Science Learning in Games*.** To fully realize the educational potential of science games, assessment of learning needs to be a component of the gaming environment, otherwise students will have only an enjoyable experience, but little or no learning will occur. Also, until assessments can be effectively embedded in games, they will not be useful as alternative ways of assessing students in the classroom or as part of accountability tests. Assessment can be used in a few ways in games. First, establishment of the learning goals during the game design phase allows the tasks of the game to be aligned with desired educational outcomes, Second, assessment can be incorporated into the game play if it used to assess how well a player is performing at key points in the game and to determine what stage of the game is educationally appropriate next for the player. In this way, game play and assessment are intertwined. Thirdly, judgment of performance against educational goals on the game overall can provide indications of the understanding that a student has gained through their game play. However, if assessment is not carefully embedded in unobtrusive ways, then the player is going to feel that the experience is neither playful nor enjoyable, and may stop playing. Evidence-centered design seems to offer a useful framework for developing assessments that can operate in the dynamic learning environments of games, but the field needs to actually use ECD to develop such assessments because doing so will help to refine the approach. Also, real examples will lead to deeper thinking about how to achieve accurate but stealthy assessments that can provide feedback to students and their teachers about the learner's progress in science knowledge and skills.

Psychometrics for Assessments in Science Simulations and Games

How Assessment is Applied in Simulations and Games. The field of educational psychometrics has grown up around the types of responses that are typical in paper-based, large-scale assessments: primarily multiple-choice and written response items. Over the years, the methods of analyzing these types of student responses have become increasingly sophisticated, progressing from Classical Test Theory to Item Response Modeling methods that can model different dimensions of students responses and even dynamically adapt the assessment to the ability of the student as the student is being assessed. In computer-based simulations and games, the ways in which a student can respond in an assessment task are greatly expanded. As a result, the old ways of describing response types as, for example, multiple-choice or written response, is too limiting and new ways of thinking about response types need to be defined. In games, particularly, any assessment task that interrupts the flow of the game can destroy the very elements of enjoyment that make it playful and engaging, so new unobtrusive measures need to be developed.

Limitations of Classical Test Theory and Item Response Theory. The complex tasks in simulations and games cannot easily be modeled using just Classical Test Theory (CTT) and Item Response Theory (IRT), the methods most commonly used in educational testing. To understand why this is the case, it is helpful to refer to the definition of "complex tasks" in computer-based testing that was proposed by Williamson et al. (Williamson, Bejar, & Mislevy, 2006) which lists four characteristics of complex tasks. The first characteristic is that the completion of the task requires the student to undergo multiple, non-trivial, domain-relevant steps and/or cognitive processes. This is true of simulations and games. For example, as shown in Figures 1-3, students in the SimScientists simulation-based assessment for ecosystems first observe a simulated ecosystem, noting the behaviors of the organisms, then construct a food web to represent their observations, and finally use a population model tool to vary the number of organisms in the ecosystem to observe outcomes over time.

The second characteristics of complex tasks is that multiple elements, or features, of each task performance are captured and considered in the determination of summaries of ability and/or diagnostic feedback (Williamson et al., 2006). The wide range of student responses and actions that are captured in simulations and games illustrates this point. These range from standard selected responses like multiple choice and short written responses to actions like drawing an arrow in a food web (Quellmalz, Timms and Buckley, in press), gathering quantitative evidence on fish, water and sediment in a lake (Squire & Jan, 2007), to the depth and angle of incision of a surgical instrument in a mannequin (Russell, 2002). In addition to response types, researchers in the field of intelligent tutoring and assessment are working on interpreting student emotions via facial recognition, skin sensors, posture, gestures and brain waves to get measures of student engagement and affect (Arroyo et al., 2009; Heraz & Frasson, 2009; Stevens, Galloway, Berka, Johnson, & Sprang, 2008) to better interpret and support student learning. Such measures will obviously require different metrics and models if and when they are incorporated into assessments.

Williamson et al. (2006) identify that the third characteristic of complex tasks is that there is a high degree of potential variability in the data vectors for each task, reflecting relatively unconstrained work product production. An example of this in the performance of complex tasks in simulations and games is when the time taken by a student to perform a task is used as a variable measurement. Unlike traditional measures that have a linear pattern in which more of something is better, time does not necessarily follow that pattern. A student who completes a task quickly might be a high performer who possesses the knowledge and skill required for the task, or he might be a low performer who does not have know and skills, but responds quickly in order to move on in the overall assessment. A student who takes a long time performing the task might be skilled in the task, but proceeds thoughtfully and carefully, or he may be unskilled and lingers because he is confused. Without being considered in conjunction with additional variables about task performance, time is not an easy variable to interpret.

The fourth characteristic in the Williamson et al. definition of complex tasks is that the evaluation of the adequacy of task solutions requires the task features to be considered as an interdependent set, for which assumptions of conditional independence typically do not hold. There is a tension between validity and reliability in assessment in simulations and games. Because of their complexity and the longer time that students spend on the tasks, simulations and games can mimic real-world scenarios and thereby provide greater validity to the assessment. At the same time, however, the use of these complex tasks reduces the number of measures that can be included in any one test and causes many of the measures to be interdependent because they are related to the same scenario, thereby reducing the reliability. While some independence can be achieved by segmenting tasks within the overall simulation or game, interdependence is hard to avoid.

The Need for an Increased Range of Psychometric Methods. Taken together, the characteristics of complex tasks in simulations and games lead to diverse sequences of tasks that produce multiple measures, often gathered simultaneously, that require metrics that are not usually found in standard assessments. The multidimensional nature of assessment in simulations and games make CTT unsuitable as a measurement method because it cannot model different dimensions of a performance simultaneously. Measurement in complex tasks involves interpreting patterns of behavior across one or more tasks. The assessments also need to be made in real time (on-the-fly) as the student is still engaged in the task, and they are often calculated based on very limited amounts of data due to the fact that the student is in the early stages of a task or series of tasks.

So, the types of measurement methods that better lend themselves to simulations and games are probability-based methods (like IRT and Bayes Nets) that can handle uncertainty about the current state of the learner, can provide immediate feedback during tasks (e.g., Model Tracing or rule-based methods like decision trees), and are able to model patterns of student behavior (e.g., Artificial Neural Networks and Bayes Nets). These methods are discussed next.

Discussion of measurement models suitable for assessment in simulations and games

The following methods have been used in simulations and games, or in related work in intelligent tutoring, and each is explained briefly and citations are given for relevant work that has used them. It is not intended to be an exhaustive list, nor a detailed explanation of the methods.

Item Response Theory – Item response models like the Rasch model have the advantage that they place estimates of student ability and item difficulty on the same linear scale, measured in logits (a log of the odds scale). This means that the difference between a student's ability estimate and the item difficulty can be used to interpret student performance. Since both the estimates of student abilities and the estimates of item difficulty are expressed in logits, they can be meaningfully compared. IRT could be a useful methodology to use in determining how much help students need in solving problems in an intelligent learning environment by measuring the gap between item difficulty and current learner ability (Timms, 2007). IRT is also useful for analyzing existing response data to obtain estimates of prior probabilities that can be used in Bayes Nets and has been used in conjunction with Artificial Neural Networks (Cooper and Stevens, 2008).

Bayes Nets - A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph. In the Bayes Net, nodes represent random variables and the edges (links between the nodes) encode the conditional dependencies between the variables. Across a series of nodes and edges a joint probability distribution can be specified over a set of discrete random variables. Figure 4 shows an example of a fragment of a Bayes Net used in the scoring of the ecosystems benchmark assessments in SimScientists. It shows how nodes in the network representing data gathered from student actions in the assessment (the lower two rows) provide information to assess the top level variables of content knowledge and science inquiry skills represented in the upper two rows.

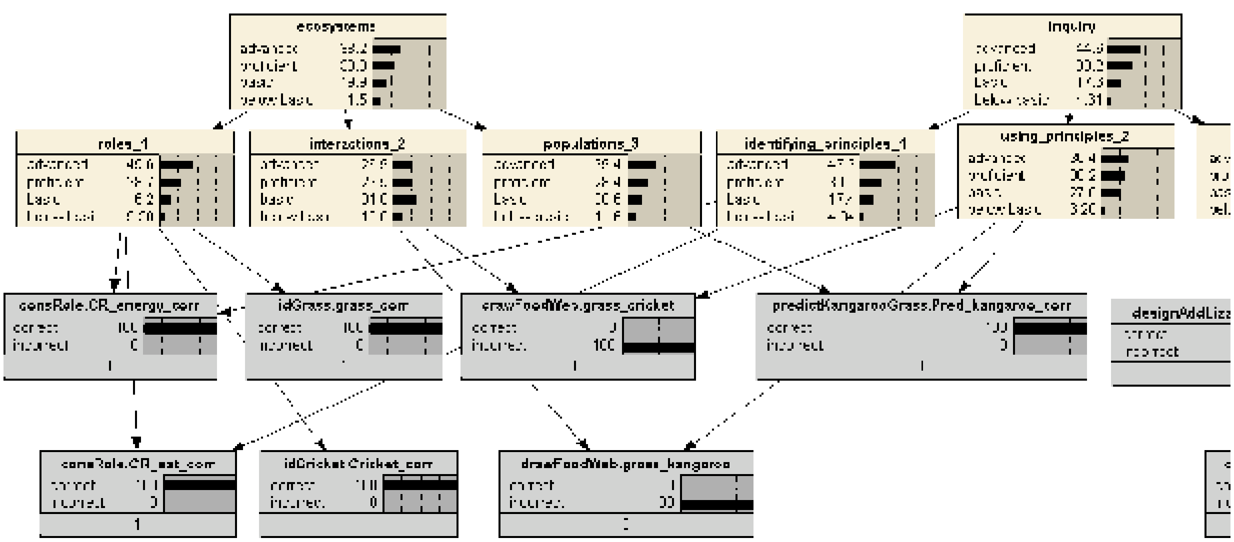

Figure 4. *Fragment of a Bayes Net From the SimScientists Ecosystems Benchmark Assessment*

Values for the edges are encoded, but not visible in this view. Data are gathered from student interactions with the simulation or game and passed to the Bayes Net where algorithms are then applied using software such as Netica to perform inference to produce estimates of probability that students possess the knowledge or skill represented via the nodes. In recent years, Bayesian networks have been widely used in intelligent tutoring systems but over the years, their use in systems for assessment has grown. Martin and VanLehn (1995) and Mislevy and Gitomer (1996) studied the applications of Bayesian networks for student assessment. Mislevy has continued this work, although not in science assessments, with Behrens in the NetPass program which assesses examinees ability to design and troubleshoot computer networks (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008). Conati et al. (2002) applied Bayesian networks to both assessing students' competence and recognizing students' intentions. WestEd is currently developing a

Bayes Net system (among other methods) in the SimScientists simulation-based science assessment.

Artificial Neural Networks – An Artificial Neural Network (ANN) is an adaptive, most often nonlinear system that learns to perform a function (an input/output map) from data. In a "supervised" ANN, the system parameters are developed in a training phase during which the system is calibrated using sample data that has already been scored. The ANN is built using a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule. After the training phase, the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (the testing phase). ANN models achieve good performance via massively parallel nets composed of non-linear computational elements, sometimes referred to as units or neurons. Each neuron has an activation level that is represented as a number and each connection between neurons also has a number, called its weight. These resemble the firing rate of a biological neuron and the strength of a synapse (connection between two neurons) in the brain. A neuron's activation depends on the activations of the neurons connected to it and the interconnection weights. Neurons are often arranged into layers. Input layer neurons have their activations set externally. ANNs have been widely used in intelligent systems, especially those in which the system needs to learn from data. In science education, an example of the use of ANNs is in the work of Stevens through a series of projects in IMMEX (Interactive MultiMedia Exercises). A recent article (Cooper & Stevens, 2008) describes the use of ANNs to assess student metacognition in problem-solving in chemistry. In addition to the ANN, IMMEX uses Hidden Markov Models to cluster a large number of performances in a predetermined number of strategies (called states) and uses IRT to model the student ability, or level of difficulty that the student has been able to reach in the problem set.

Model Tracing – the model tracing approach was developed for the cognitive tutors produced by the Pittsburgh Advanced Cognitive Tutors (PACT) center at Carnegie Mellon University. Model tracing works by comparing the student's solution of a problem to an expert system for the domain of interest. Production rules, or rules about knowledge and skills in a given domain are, in this system, based on an approach from the work of cognitive scientist John Anderson's ACT-R model representing skill-based knowledge (Anderson, 1993; Anderson & Lebiere, 1998). As the student progresses through the problem solving, the model tracing system generates at each step the set of all possible next steps by referring to the production rules. These possible "next steps" are not displayed to students but are used by the computer to evaluate the quality of the student's next step in problem solving. The computer-generated set of possible steps is called the *conflict set*, and the decision as to which is the best next step to take from the entire set of possible steps is called *resolution of the conflict set*. The computer assesses each of the possible next steps in the conflict set and decides if it is productive, counter-productive or illegal (one that violates a fundamental principle). It is the group of productive solutions which the tutor then evaluates as to which is most teachable and presents those options to the student.

Rule-based methods – For immediate, formative assessment of student actions, rule-based methods that are simpler than the other methods discussed above, can be appropriate. Rule based methods are ones that employ some logic method to decide how to interpret a student action. A simple example would be posing a multiple-choice question in which the distractors (wrong answer choices) were derived from known misconceptions in the content being assessed. The student's incorrect response could then be diagnosed and immediate action can be taken, such as providing coaching. This type of diagnosis is the basis of the work of Minstrell and Kraus in their work with the DIAGNOSER software that assesses students in science and diagnoses their understanding and misconceptions (Minstrell & Kraus, 2007). An example of a more complex rule-based method is the decision tree, which applies a logic chain to categorize the student response. The logic chain is represented in a diagram with logic gates that direct the program to implement one of two choices depending on whether the logic test is true or false. An example from the SimScientists project is shown in Figure 5, which shows the logic behind the generation of feedback messages to students as they develop a food web diagram to represent the interactions of organisms that they observed in the mountain lake ecosystem as illustrated in Figure 3.

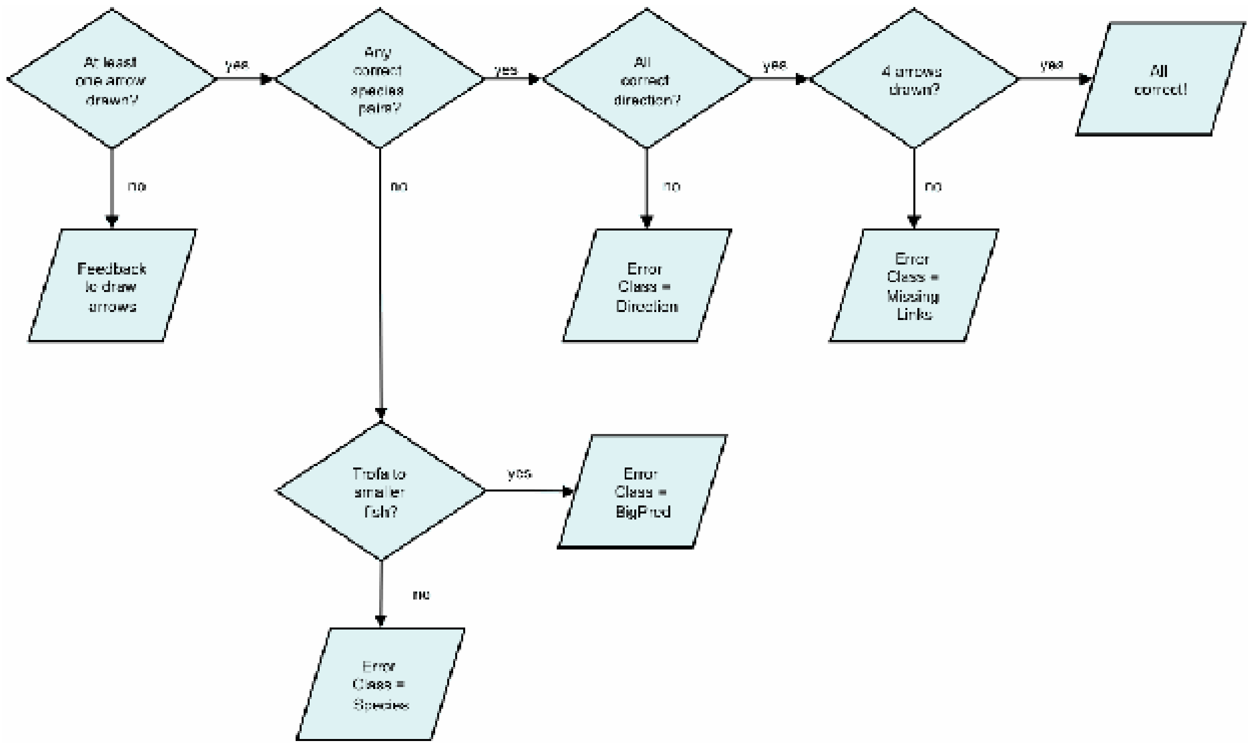

Figure 5. *Example of a Decision Tree for Diagnosing Student Misconceptions in the SimScientists Ecosystems Embedded Assessment*

Summary and Recommendations for Research

The use of simulations *for* and *of* learning will benefit from studies of effective and efficient assessment designs. Pilot projects can provide evidence of the quality, utility and feasibility of simulation-based assessments. Research on cognitively-based principles for designing simulation-based assessments can both test hypotheses about the benefits of static, active and interactive modalities for testing types of science knowledge (declarative, schematic, procedural, schematic) and contribute to design principles to be used by the assessment community. Limitations and affordances of science simulations for English learners and students with disabilities are just beginning to be explored (Silberglitt, submitted; Kopriva, et al., 2009). Learning benefits of science simulations for complex learning in post secondary courses, informal science education environments (museums, after school programs), and practical settings (health clinics, wellness programs) should be explored. All such investigations call for research-based, rigorously designed assessments that meet their intended purposes, such as diagnostic assessment or summative assessment. All such research calls for documentation of the assessment designs and technical quality.

Much more research is needed on how assessment can be built into appropriate serious games in science to demonstrate that learning can be assessed in reliable and valid ways. Until it has been established that assessment can be built into science games, the field cannot move on to address the question of how games can be used in formative assessments in the classroom or, perhaps, in accountability assessments to provide evidence of complex skills that cannot be assessed by existing formats.

The growing use of simulations and games to assess and promote student learning requires an expansion of the psychometric methods that are used to measure and interpret student performance. Further research is needed to investigate what are the most effective methods to assess learning in complex tasks in games and simulations. This will require that we expand our range of psychometric tools to include such things as Bayesian Networks, Artificial Neural Networks, Model Tracing, Rule-based methods and possibly even other methods to take full advantage of these new assessment media. Education researchers will need to work with other disciplines like computer science where expertise on pattern analysis of complex data already exists. Also, training for future psychometricians should include learning about and using this expanded methodological toolkit.

REFERENCES

Adams, W. K., Reid, S., LeMaster, R., McKagan, S., Perkins, K., Dubson, M., et al. (2008). A study of educational simulations part II—Interface design. *Journal of Interactive Learning Research, 19*(4), 551-577.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). *Emotion Sensors Go To School.* Paper presented at the Artificial Intelligence in Education, Brighton, UK.

Barab, S. (2006). Design-based research: a methodological toolkit for the learning scientist. In S. K (Ed.), *The handbook of the learning sciences* (pp. 153–170). Cambridge: Cambridge University Press.

Barab, S., & Dede, C. (2007a). Games and immersive participatory simulations for science education: an emerging type of curricula. *Journal of Science Education and Technology,* 16(1), 1–3.

Barab, S., Sadler, T., Heiselt, C., Hickey, D., & Zuiker, S. (2007b). Relating narrative, inquiry, and inscriptions: a framework for socioscientific inquiry *Journal of Science Education and Technology,* 16.

Behrens, J. T., Frezzo, D., Mislevy, R., Kroopnick, M., & Wise, D. (2008). *Structural, Functional, and Semiotic Symmetries in Simulation-Based Games and Assessments.* In E. Baker, J. Dickieson, W. Wufleck & H. F. O'Neil (Eds.), Assessment of Problem Solving Using Simulations (pp. 59-80). New York: Lawrence Erlbaum Associates.

Beal, C., Beck, J., Westbrook, D., and Cohen, P. (2002). Intelligent modeling of the User in Interactive entertainment. AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment.

Boyle, A. (2008). *Sophisticated tasks in e-assessment: What are they and what are their benefits?* Paper presented at 2005 International Computer Assisted Assessment Conference, Loughborough University, United Kingdom. Retrieved on July 17, 2008 from http://www.caaconference.com/pastConferences/2005/proceedings/BoyleA2.pdf

Brown, J., Hinze, S., Pellegrino, J. W. (2008). Technology & formative assessment. In T. Good (Ed.) *21st Century Education.* Thousand Oaks, CA: Sage.

Buckley, B. C., Gobert, J., Horwitz, P., & O'Dwyer, L. (in press, 2009). Looking inside the black box: Assessing model-based learning and inquiry in BioLogica *International Journal of Learning Technologies*.

Buckley, B.C., Quellmalz, E.S., & Davenport, J. (submitted). SimScientists: Theoretical Foundations.

Buckley, B. C., Gobert, J., Kindfield, A. C. H., Horwitz, P., Tinker, B., Gerlits, B., et al. (2004). Model-based teaching and learning with hypermodels: What do they learn? How do they learn? How do we know? *Journal of Science, Education and Technology, 13*(1), 23-41.

Cooper, M. M., & Stevens, R. (2008). Reliable multi-method assessment of metacognition use in chemistry problem solving. *Chemistry Education Research and Practice*, 9, 18-24.

Cross, T. R., & Cross, V. E. (2004). Scalpel OR mouse? *The American Biology Teachers College Record, 66*(6), 408-411.

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 180.

Doerr, H. (1996). Integrating the study of trigonometry, vectors, and force through modeling. *School Science and Mathematics, 96*, 407-418.

Gee, J. P. (2005). Learning by design: Good video games as learning machines *E-Learning, 2(1)*, 5–16.

Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. *International Journal of Science Education, 22*(9), 891-894.

Heraz, A., & Frasson, C. (2009). *Predicting Learner Answers Correctness through Brainwaves Assessment and Emotional Dimensions*. Paper presented at the Artificial Intelligence in Education, Brighton, UK.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher, 30*, 141-158.

Hickey, D. T., Ingram-Goble, A. A., & Jameson, E. M. (2009). Designing Assessments and Assessing Designs in Virtual Educational Environments. *Journal of Science Education and Technology*, 18, 187–208.

Hmelo-Silver, C.E., Jordan, R., Liu, L. Gray, S., Demeter, M., Rugaber, S., Vattan, S., & Goel, A. (2008). Focusing on function: Thinking below the surface of complex science systems. *Science Scope.*

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. T. (2003). Integrating curriculum, instruction, assessment, and evaluation in a technology-supported genetics learning environment. *American Educational Research Journal, 40*(2), 495-538.

Horwitz, P., Gobert, J., Buckley, B. C., & Wilensky, U. (2007). *Modeling across the curriculum: Annual report to NSF*. Concord, MA: The Concord Consortium.

Jackson, S. L., Stratford, S. J., Krajcik, J., Soloway, E. (1996). A learner-centered tool for students building models. *Communications of the ACM, 39*(4), 48-49.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2008). Studying Situated Learning in a Multiuser Virtual Environment. In E. Baker, J. Dickieson, W. Wufleck & H. F. O'Neil (Eds.), *Assessment of Problem Solving Using Simulations.* New York: Lawrence Erlbaum Associates.

Ketelhut, D. J., Dede, C., Clarke, J., & Nelson, B. (2006). A multi-user virtual environment for building higher order inquiry skills in science. Paper presented at the American Educational Research Association, San Francisco, CA.

Kopriva, R., Gabel, D., Bauman, J. (2009). *Building comparable computer-based science items for English Learners: Results and insights from the ONPAR project.* National Conference on Student Assessment (NCSA), Los Angeles, CA.

Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000, April). *Inquiry-based science supported by technology: Achievement and motivation among urban middle school students.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Land, S. M. (2000). Cognitive requirements for learning with open-ended learning environments. *Educational Technology, Research and Development, 48*(3).

Ma, J., & Nickerson, J. V. (2006). Hands-on, simulated, and remote laboratories: A comparative literature review. *ACM Computing Surveys, 38*(3), 1-24.

Magnussen, R. (2005). Learning games as a platform for simulated science practice. Proceedings of DiGRA 2005 Conference: Changing Views – Worlds in Play.

McQuiggan, S. W., Goth, J., Ha, E., Rowe, J. P., & Lester, J. C. (2008). *Student Note-Taking in Narrative-Centered Learning Environments: Individual Differences and Learning Effects.* Paper presented at the Proceedings of the Ninth International Conference on Intelligent Tutoring Systems, Montreal, Canada.

McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2008). *Affective Transitions in Narrative-Centered Learning Environments.* Paper presented at the Proceedings of the Ninth International Conference on Intelligent Tutoring Systems, Montreal, Canada.

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

Minstrell, J., & Kraus, P. A. (2007). *Applied Research on Implementing Diagnostic Instructional Tools*. FACET Innovations.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H.Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (437-446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., Hafter, A., Hamel, L., Kennedy, C., Long, K., Morrison, A. L., Murphy, R., Pena, P., Quellmalz, E., Rosenquist, A., Songer, N., Schank, P., Wenk, A., & Wilson, M. (2003) *Design patterns for assessing science inquiry* (PADI Technical Report 1) Menlo Park, CA: SRI International, Center for Technology in Learning.

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

Neulight, N., Kafai, Y., Kao, L., Foley, B., & Galas, C. (2007). Children's participation in a virtual epidemic in the science classroom: Making connections to natural infectious diseases. *Journal of Science Education and Technology, 16*(1), 47-58.

Norman, D.A. (1993). *Things that makes us smart.* Reading, MA; Adddison-Wesley.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Prensky, M. (2001). *Digital Game-Based Learning*. McGraw-Hill.

Quellmalz, E. S., DeBarger, A., Haertel, G., & Kreikemeier, P. (2005). *Validities of science inuqiry assessments: Final report*. Menlo Park, CA: SRI International.

Quellmalz, E. S., DeBarger, A. H., Haertel, G., Schank, P., Buckley, B., Gobert, J., Horwitz, P., & Ayala, C. (2008). Exploring the role of technology-based simulations in science assessment: The Calipers Project. In *Science assessment: Research and practical approaches.* Washington, DC: NSTA.

Quellmalz, E. S., & Haertel, G. (2004, May). *Technology supports for state science assessment systems.* Paper commissioned by the National Research Council Committee on Test Design for K-12 Science Achievement.

Quellmalz, E. S., & Haertel, G. D. (2008). Assessing new literacies in science and mathematics. In D. J. Leu, Jr., J. Coiro, M. Knowbel, & C. Lankshear (Eds.) *Handbook of research on new literacies.* Mahwah, NJ: Erlbaum.

Quellmalz, E.S. & Pellegrino, J.W. (2009). Technology and testing *Science,* 323, 75-79.

Quellmalz, E.S., Timms, M.J., & Buckley, B.C. (in press). The promise of simulation-based science assessment: The calipers project. *International Journal of Learning Technologies.*

Reiber, L. P., Tzeng, S., & Tribble, K. (2004). Discovery learning, representation, and explanation within a computer-based simulation. *Computers and Education, 27*(1), 45–58.

Reif, R. J. (2004). *Virtual anatomy alternatives*. Santa Fe, New Mexico: New Mexico Public Education Department.

Repenning, A., & Ioannidou, A. (2005). *Mr. Vetro: A collective simulation framework.* Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005, Montreal, Canada.

Robison, J., McQuiggan, S., & Lester, J. (2009). *Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems.* Paper presented at the Proceedings of the International Conference on Affective Computing & Intelligent Interaction, Amsterdam, Netherlands.

Rowe, J. P., McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2009). *Off-Task Behavior in Narrative-Centered Learning Environments.* Paper presented at the Artificial Intelligence in Education, Brighton, UK.

Russell, M. (2002). *How Computer-Based Technology Can Disrupt the Technology of Testing and Assessment.* Washington, DC: Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education.

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

Scalise, K., Bernbaum, D. J., Timms, M., Harrell, S. V., Burmester, K., Kennedy, C.A. & Wilson, M. (2007). Adaptive technology for e-learning: Principles and case studies of an emerging field. *Journal of the American Society for Information Science and Technology, 58* (14), 2295-2309.

Scalise, K., Timms, M., Clark, L., & Moorjani, A. (2009). *Student learning in science simulations: What makes a difference*. Paper presented at the Conversation, Argumentation, and Engagement and Science Learning, American Educational Research Association, San Diego, CA.

Schwartz, D. L., & Heiser, J. (2006). Spatial representations and imagery in learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*. Cambridge: Cambridge University Press.

Shaffer, D.W., Hatfield, D., Svarovsky, G.N., Nash, P., Nulty, A., Bagley, E., et al. (2009). Epistemic network analysis: A prototype for 21st century assessment of learning. *International Journal of Learning and Media, 1(*2), 33-53.

Shute, V. J., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., et al. (2009). Modeling, Assessing, and Supporting Key Competencies Within Game Environments. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge.* New York: Springer-Verlag.

Shute, V.J., Ventura, M., Bauer, M., and Zapata-Rivera, D., (2008). *Monitoring and fostering learning through games and embedded assessments.* Research Report (RR-08-69). Princeton, NJ: Educational Testing Service.

Silberglitt, M. (submitted). Integrating Simulation-Based Science Assessments into Balanced State Science Assessment Systems.

Squire, K. D., & Jan, M. (2007). Mad City Mystery: Developing Scientific Argumentation Skills with a Place-based Augmented Reality Game on Handheld Computers. *Journal of Science Education and Technology* 16(1), 5-29.

Srinivasan, S., & Crooks, S. (2005). Multimedia in a Science Learning Environment. *Journal of Educational Multimedia and Hypermedia, 14*(2), 151-167.

Stewart, I., & Golubitsky, M. (1992). *Fearful symmetry: Is God a geometer?* Massachusetts: Blackwell Cambridge.

Stieff, M., & Wilensky, U. (2003). Connected chemistry – incorporating interactive simulations into the chemistry classroom. *Journal of Science Education and Technology,* 12(3), 285-302.

Stevens, R. H., Galloway, T. L., Berka, C., Johnson, R., & Sprang, M. (2008). *Assessing Student's Mental Representations of Complex Problem Spaces with EEG Technologies* Paper presented at the Human Factors and Ergonomics Society 52nd Annual Meeting, New York, NY.

Timms, M. (2007). Using item response theory (IRT) in an intelligent tutoring system. Proceedings of the Artificial Intelligence in Education 2007 Conference, Marina Del Ray, CA. IOS Press, Washington DC. *Frontiers in Artificial Intelligence and Applications. Vol 158 (213-221).*

Vendlinski, T. & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *Journal of Technology, Learning, and Assessment, 1*(3). Available from http://www.jtla.org.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3-118.

Wideman, H. H., Owston, R. D., Brown, C., Kushniruk, A., Ho, F., & Pitts, K. C. (2007). Unpacking the potential of educational gaming: A new tool for gaming research. *Simulation Gaming, 38*(1), 10-30.

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). *Automated Scoring of Complex Tasks in Computer-Based Testing*. Mahwah, NJ: Lawrence Erlbaum Associates.