CHAPTER 10

Exploring the Role of Technology-Based Simulations in Science Assessment: The Calipers Project

Edys S. Quellmalz
WestEd

Angela H. DeBarger, Geneva Haertel, and Patricia Schank
SRI International

Barbara C. Buckley, Janice Gobert, and Paul Horwitz
Concord Consortium

Carlos C. Ayala
Sonoma State University

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator, and transferred to WestEd under Grant No 0741029. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Contact: *equellm@wested.org*. Current versions of the assessments can be found at *http://wested.org/calipers*.

Requirements of the No Child Left Behind (NCLB) law for science testing at the elementary, middle, and secondary levels by the 2007–2008 school year are renewing scrutiny of available assessments as evidence of what students should know and be able to do in science. It is widely recognized that there is a major disconnect between the content and structure of large-scale accountability tests and classroom formative assessment practices. Traditional tests tend not to be aligned with the challenging goals set forth in the National Science Education Standards (NRC 1996) and state science standards and are too limited to capture the deep conceptual understandings at the heart of science reform.

A paramount lesson learned from earlier reform efforts is the need to align key components of the educational system—i.e., standards, curricula, and assessments (Quellmalz, Shields, and Knapp 1995; Smith and O'Day 1991). Effective reform programs promote student achievement by placing a greater emphasis on conceptual understanding and application to everyday situations, increasing use of technologies, and developing new forms of assessment (Shields, Marsh, and Adelman 1998). Several decades of science learning now inform our understandings of how to design assessments to probe what students know and can do (Bransford, Brown, and Cocking 2000; Pellegrino, Chudowsky, and Glaser 2001). Central to research-based test design is a shift from questions on discrete, factual content to questions that focus on relationships among concepts and tasks that require integration of reasoning and inquiry within significant, recurring, extended academic and applied problems. However, many assessments continue to rely on discrete items, primarily using the multiple-choice format. These tests still favor shallow content coverage (Quellmalz and Haydel 2002). Clearly, better methods for capturing compelling evidence of student science learning, both content knowledge and inquiry skills, must be made available.

The powerful capabilities of technology hold the key to transforming current assessment practices at both the state and classroom levels (Quellmalz and Haertel 2004). Currently, external, technology-based accountability assessments do not incorporate complex performance tasks, nor do technology-rich curricula yet employ principled assessment designs that provide student performance data that meet the standards of technical quality required for external assessments. What is needed, therefore, is development of assessment designs and examples that can take advantage of technology to bring high-quality assessments of complex performances into science tests with accountability goals and with formative goals. In this chapter, we describe a project funded by the National Science Foundation, "Calipers: Using Simulations to Assess Complex Science Learning."

Value and Uses of Simulations in Education

Increasingly, simulations are playing an important role in science and mathematics education. Simulations support conceptual development by allowing students to explore relationships among variables in models of a system. Simulations can facilitate knowledge integration and a deeper understanding of complex topics, such as genetics, environmental science, and physics (Buckley et al. 2004; Hickey et al. 2003; Krajcik et al. 2000; Doerr 1996). Moreover, simulations have the potential to represent content and relationships in ways that can reduce reading demands and allow students to "see" a variety of concepts and relationships (e.g., pictures, graphs, tables). Simulations are well-suited to investigations of interactions among multiple variables in models of complex systems (e.g., ecosystems, weather systems, wave interactions) and to experiments with dynamic interactions exploring spatial and causal relationships. Technology allows students to manipulate an array of variables, observe the impact, and try again. The technology can provide immediate feedback. Simulations also can make available realistic problem scenarios that are difficult or impossible to create in typical classrooms.

Simulations can allow students to engage in the kinds of investigations that are familiar components of hands-on curricula, but also to explore problems and discover solutions they might not be able to investigate in classrooms. They also allow experimentation with phenomena that are too large or small, too fast or slow, or too expensive or dangerous. In addition, simulations do not require the logistical planning involved in setting up equipment for hands-on science experiments.

Numerous studies have discussed the benefits of using simulations to support student learning. Model-It has been used in a large number of classrooms, and positive learning outcomes based on pretest-posttest data have been 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. 1995). After participating in the Connected Chemistry project, which uses 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 and 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 and Frederiksen 1998). An implementation study of the use of BioLogica by students in eight high schools (Buckley et al. 2004) showed an increase in genetics content knowledge in specific areas, as well as an increase in genetics problem-solving skills. Studies conducted with BioLogica suggest that the activities maintain student engagement while also linking their explorations to underlying content in genetics (Horwitz and Christie 1999).

Calipers Project Goals

The Calipers project is a two-year demonstration project that aims to use technology-supported "benchmark assessments" to bridge the gap between external summative assessments of principled design and high technical quality and curriculum-embedded formative assessments.

The Calipers project has developed a new generation of technology-based science assessments that measure student science knowledge of the relationship of multiple components in a system and inquiry skills integrated throughout extended problem-based tasks. The Calipers simulation-based assessments are intended to augment available assessment formats; make high-quality assessments of complex thinking and inquiry accessible for classroom, district, program, and state testing; and reduce economic and logistical barriers that impede the use of rich science assessment. The Calipers project provides evidence of the feasibility, usability, and technical quality of the new simulation-based assessments. In addition, the project has prepared a plan for development of a larger pool of simulation-based complex assessments linked to key strands in the AAAS *Atlas of Science Literacy* (AAAS 2001) and core National Science Education Standards.

Development of the Calipers Assessments

The development of the Calipers assessments includes a principled approach to the assessment design, alignment of the assessments with key science standards and representative science curricula, pilot testing and revisions, and a plan for development of additional environments and assessments.

The Calipers assessments were designed to test science knowledge and inquiry strategies in two fundamental life and physical science areas. Life science standards related to populations and ecosystems were chosen for one of the simulation prototypes. Physical science standards related to forces and motion were selected for the second set of prototypes. For each area, Concord Consortium designed the model of the environment to be simulated, and SRI International and WestEd designed the assessment items and tasks related to the environments.

Design of the assessments followed a principled assessment design approach (Mislevy et al. 2003). The science knowledge and inquiry skills to be tested were specified. The evidence that would provide observations of achievement of the knowledge and inquiry was specified in terms of the types of student responses to be elicited and scoring criteria. Features of tasks and items that would elicit evidence of achievement were specified. For both of the content areas, the science knowledge and inquiry abilities were aligned with the AAAS Benchmarks and key ideas and the National Science Education Standards.

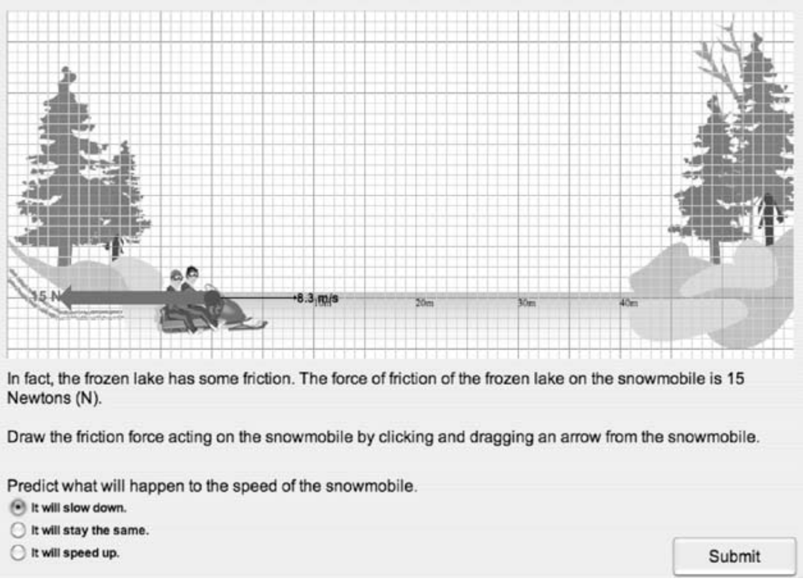
Guidelines for the Calipers assessment tasks included (1) specification of a driving, authentic problem, (2) design of items and tasks to take advantage of the simulation technology, (3) alignment with standards, and (4) alignment with the types of problems and activities presented in curricula.

Simulation-Based Assessments for Forces and Motion

The setting selected to simulate principles of force and motion included skiers and snowmobiles on a mountain. The driving problem was the need for a student dispatcher to coordinate the rescue of injured skiers by snowmobile units. The simulation engine developed by Concord Consortium built on its existing Dynamica engine. To demonstrate the flexibility of the environment for assessments at a range of levels of complexity, three assessments were developed to test concepts and inquiry strategies appropriate from the early middle school grades to grade 9 physical science. Students were asked to predict and explain what would happen to the snowmobile on varying terrain (e.g., sloped, frictionless). Student manipulations of the simulation included drawing force arrows and running the simulation.

Figure 10.1 presents a screen shot of a scene within one of the Mountain Rescue assessments. Students are asked to draw an arrow (see lower left) depicting the magnitude and direction of the friction force acting on the snowmobile and predict what will happen to the snowmobile. In a subsequent screen, after running the simulation to see if their prediction was correct, students are asked to explain to the rescue team why the snowmobile behaved as it did. Student manipulations of the simulation and responses to the question provide evidence of their knowledge of balanced and unbalanced forces on surfaces with and without friction. Other tasks and scenarios test inquiry skills for prediction, explanation, and interpretation of graphs. Questions related to simpler and more complex knowledge are asked in the three separate assessments, and additional inquiry skills, such as designing the experiment and communicating recommendations, are tested.

Figure 10.1 Force and Motion Assessment 1: Friction Force Drawing and Prediction Items



In fact, the frozen lake has some friction. The force of friction of the frozen lake on the snowmobile is 15 Newtons (N).

Draw the friction force acting on the snowmobile by clicking and dragging an arrow from the snowmobile.

Predict what will happen to the speed of the snowmobile.

● It will slow down.
○ It will stay the same.
○ It will speed up.

As students participate in the force and motion assessments, the computer captures their answers to questions whether in the form of multiple choice, short answer, or essay. The computer records the magnitude and direction of arrows they draw and captures their manipulations of the simulations. When students experiment with the snowmobile speed to determine the best speed for getting to skiers on an icy hill, the computer captures the speed selected for each experimental trial. This information can be used to examine how each student in an entire class performs an experiment, a task that cannot be done in a classroom laboratory. We can determine if students have chosen experimental values that cover the range necessary and if they were systematic in exploring the range of values. Finally, we can determine if they were successful in accomplishing the task.

For many types of responses (e.g., multiple choice, drawing force arrows), the computer can automatically produce a score based on a rubric created by the Calipers assessment developers. To score multiple-choice questions, the computer identifies whether students selected the correct answer. Students' responses also can be automatically coded to facilitate the diagnosis of problem-solving strategies and types of errors in understanding. For example, in the first force and motion assessment students are asked to calculate how long it will take to travel a certain distance at a given speed. Students first select the correct formula for performing this calculation, then enter the values for distance and speed. The computer calculates the answer and students are asked to evaluate their answer. The computer automatically scores student responses using a rubric that awards 2 points for selecting the correct formula the first time, 1 point for selecting it on the second or third try, and 0 points for failing to select the correct formula within three tries. A similar scheme awards points for entering the correct values into the equation. If students accurately evaluate their answers, another point is awarded. In contrast to assessments that score only the final answer, this enables us to pinpoint where students have difficulty.

When students are conducting experiments to determine the best speed for the snowmobile to use to reach the skiers on the icy hill, the score is determined by examining if each experimental value entered is closer to or further away from the "correct" speed. Students receive one point for moving closer to the target speed. For the entire task, we average all the runs that a student makes. In addition, we take into account whether they bracket the target speed and whether they repeat any trials. Optimum performance on the experiment would include one run with a speed less than target but greater than the start speed of the team that failed at the task, one run with a speed greater than the target, and one at target.

For the constructed-response text-based questions, the computer captures the text exactly as the student types it. Another program displays the answers of the entire class, along with the question and the scoring rubric. The teacher or researcher reads the response, compares it to the rubric, and enters a score that the computer captures and integrates into the students' records.

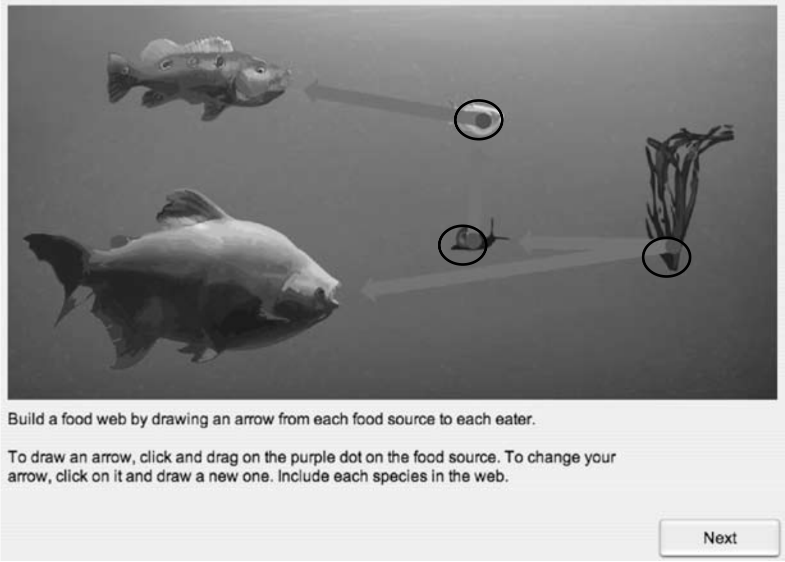
When all of the responses have been scored by computer and humans, the results are placed in a database that can be explored in a variety of ways. A teacher or researcher can see how well students are performing on specific content or inquiry targets or how well students are performing on the assessment as a whole. Researchers can also compare how well students who are working with different curricula perform.

Simulation-Based Assessments for Ecosystems

The setting selected to simulate principles for populations and ecosystems is a newly discovered lake in the jungle. The driving problem is to explore the lake and describe its ecosystem. The simulation engine for modeling the ecosystem has been developed by Concord Consortium, building on its existing Biologica engine. To demonstrate the flexibility of the environment for assessments at a range of levels of complexity, three assessments were developed to test concepts and inquiry strategies appropriate from the early middle school grades to high school biology. Students are asked to identify the relationships of the fish and plant species and predict and explain the effects of introducing new fish species. Manipulations of the simulation include drawing food webs and varying the number of predator and prey.

Figure 10.2 presents a screen shot of a scene within one of the Fish World assessments, in which students observe species and draw a food web. Figure 10.3, p. 200, presents a screen shot of the population level of the ecosystem.

Figure 10.2 Analyzing the Relationships Among Organisms in the Ecosystem



Build a food web by drawing an arrow from each food source to each eater.

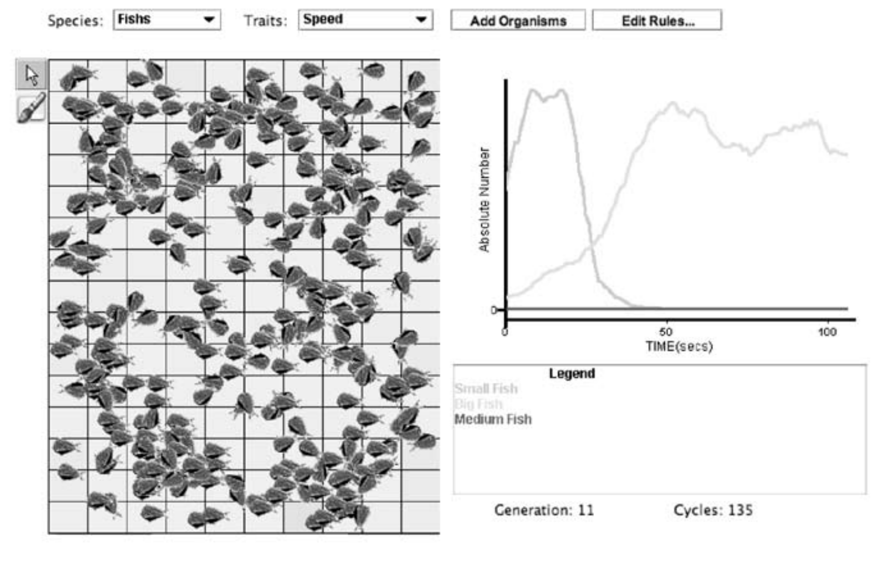
To draw an arrow, click and drag on the purple dot on the food source. To change your arrow, click on it and draw a new one. Include each species in the web.

Note: In the color version of this screen shot, viewers see three purple dots. They are circled here.

Several tasks have been designed using this layout—students modify variables and values that determine the size of populations of different organisms in the ecosystem over time. Changing population sizes are shown in a simulation and in a dynamically generated line graph.

As in the force and motion assessments, students' answers to the explicit questions and their actions manipulating the simulation are recorded by the computer and scored either automatically or by human scorers. The scores can be displayed by concept and inquiry skill, providing teachers and districts with standards-based feedback on the benchmark assessment. If the assessments were to be used for accountability, structured rater training and scoring sessions would produce interrater reliability data for the constructed-response items.

Figure 10.3 Population Dynamics in Fish World



Species: Fishs

Traits: Speed

Generation: 11 Cycles: 135

Technical Quality of the Calipers Assessments

The Calipers assessments were first tested with small numbers of students for the feasibility of the navigation and questions. The assessments were then pilot tested in classrooms. Classes were selected that had completed units addressing ecosystems or force and motion. The classes varied in their prior use of technology. For the ecosystem assessments, students completed both a Calipers simulation-based assessment and a set of items developed for item clusters on the same content and inquiry by the AAAS project.

At the time of preparation of this chapter, the Calipers project had collected data from a variety of sources to document the assessments' technical quality. Reviews by external experts of alignment of the Calipers assessments with national science standards as well as the quality of the science content and items contributed evidence of the validity of the items. Cognitive analyses of students thinking aloud as they responded to the items contributed evidence of construct validity. Analyses of data from the pilot testing of the force and motion assessments indicated that the items discriminated between high and low science achievers and seemed to be working well as indicated by the spread of responses and the fit of the items to the IRT model, meaning that they were all contributing to the measurement of the force and motion content being tested. Similar data analyses are currently underway for the ecosystem assessments.

Promise of Simulation-Based Science Assessments

The Calipers demonstration project aimed to provide evidence of the feasibility, technical quality, and utility of simulation-based science assessments. The scientifically based principles underlying the simulation environments can be re-used for both assessment and instruction. For example, the ecosystem environment can be adapted for other aquatic (e.g., saltwater) or terrestrial (e.g., Arctic) biomes. The simulations can be used to design items testing factual content as well as interrelated knowledge of systems. Inquiry tasks asking students to design, conduct, analyze and interpret data, and communicate findings can be developed. Simulation environments developed for fundamental science systems can be re-used for elementary, middle, and secondary levels. Tasks and items developed in relation to the environments can be developed for curriculum-embedded and formative assessment activities or for external accountability. Reports linking students' scores to content and inquiry standards can provide valuable information about student progress. Most important, simulations can permit assessment of knowledge and standards not well measured by paper-based formats. The development of systematically designed science simulations promises to revolutionize both instruction and assessment.

References

American Association for the Advancement of Science (AAAS). 2001. *Atlas of science literacy*. Washington, DC: American Association for the Advancement of Science.

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

Buckley, B. C., J. D. Gobert, A. C. H. Kindfield, P. Horwitz, R. F. Tinker, B. Gerlits, U. Wilensky, C. Dede, and J. Willett. 2004. Model-based teaching and learning with BioLogica™: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology* 13: 23–41.

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

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

Horwitz, P., and M. Christie. 1999. Hypermodels: Embedding curriculum and assessment in computer-based manipulatives. *Journal of Education* 181: 1–23.

Jackson, S., S. Stratford, J. Krajcik, and E. Soloway. 1995. Model-It: A case study of learner-centered software for supporting model building. Paper presented at the Working Conference on Technology Applications in the Science Classroom, Columbus, OH.

Krajcik, J., R. Marx, P. Blumenfeld, E. Soloway, and B. Fishman. 2000. 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 (April).

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

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

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

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

Quellmalz, E. S., and A. M. Haydel. 2002. Using cognitive analysis to study the validities of science inquiry assessments. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Quellmalz, E. S., P. Shields, and M. Knapp. 1995. *School-based reform: Lessons from a national study*. Washington, DC: U.S. Government Printing Office.

Shields, P. M., J. A. Marsh, and N. E. Adelman. 1998. *Evaluation of NSF's Statewide Systemic Initiatives (SSI) program: First year report*. Menlo Park, CA: SRI International.

Smith, M. S., and J. O'Day. 1991. Systemic school reform. In S. Fuhrman and B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* 33: 93–108.

Stieff, M., and U. Wilensky. 2003. Connected Chemistry—Incorporating interactive simulations into the chemistry classroom. *Journal of Science Education and Technology* 12: 285–302.

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