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 Ayala
Sonoma State University

Introduction

The powerful capabilities of technology hold the key to transforming current assessment practice at both the state and classroom levels by changing the range of student outcomes tested, how, where and when testing takes place, and the evidence available for understanding student learning. (Quellmalz & Haertel, 2004). Technologies can present students with rich task environments that model systems in the natural world. In particular, simulations can present authentic environments structured according to principles in the domain. They can be used to test students' knowledge of structural components and relationships. Because simulations are dynamic and can be manipulated., students can demonstrate their abilities to engage in active inquiry. Therefore, technologies can elicit, collect, document, analyze, appraise and display kinds of student performances that have not been readily accessible through traditional testing methods. As online testing increases, computer simulations are being seen as a potentially more affordable option for both classroom and large-scale administration of performance assessments. Simulations can offer several advantages over hands-on tasks in terms of costs (Baxter, 1995). With simulations, administration to a large sample of students simultaneously is possible, and records of student performance can be captured easily.

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. In this paper, we describe a project funded by the National Science Foundation, "Calipers: Using Simulations to Assess Complex Science Learning," which is developing assessment designs and prototypes that can take advantage of technology to bring high-quality assessments of complex performances into science tests with either accountability or formative goals.

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

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

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). These affordances of simulations may permit students from diverse language backgrounds and learning styles to understand the demands of assessment tasks and questions and to also have alternative ways to show what they know and can do. 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. Importantly, 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, and 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, 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.

Research on Simulations and Student Learning

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 & 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). Studies conducted with BioLogica suggest that the activities maintain student engagement while also linking their explorations to underlying content in genetics (Horwitz & Christie, 1999).

The Calipers Project Goals

The Calipers project is a two-year demonstration project that aims to use principled design to develop technology-supported "benchmark assessments" with technical quality to bridge the gap between external summative assessments and curriculum-embedded formative assessments. The Calipers project is developing a new generation of technology-based science assessments that will measure student science knowledge of the relationship of multiple

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

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 is documenting the feasibility, usability, and technical quality of the new simulation-based assessments. In addition, a project goal is to prepare a plan for development of a larger pool of simulation-based complex assessments linked to key strands in the AAAS *Atlas of Science Literacy* and core National Science Education Standards (NSES).

Development of the Calipers Assessments

The Calipers assessments were developed according to methods recommended by research and standards for test development (AERA/APA/NCME, 1999; Pellegrino, et. al, 2002; Quellmalz, et. al, 2005). The Calipers assessments were shaped by a principled approach to the assessment design, aligned with key science standards and representative science curricula, pilot tested, and revised. The Calipers assessments have been designed to test science knowledge and inquiry strategies in two fundamental life and physical science areas found in both the NSES and the AAAS *Benchmarks for Science Literacy*. 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.

Design of the assessments followed a principled assessment design framework that linked the knowledge and skills to be tested (student model), to features of tasks in which students can demonstrate the knowledge and skills (task model), to evaluations of student proficiency (evidence model) (Mislevy, et al., 2003). The evidence model that would provide observations of achievement of students' knowledge and inquiry was specified in terms of the types of student responses to be elicited and the scoring criteria. Features of tasks and items that would elicit evidence of achievement were specified.

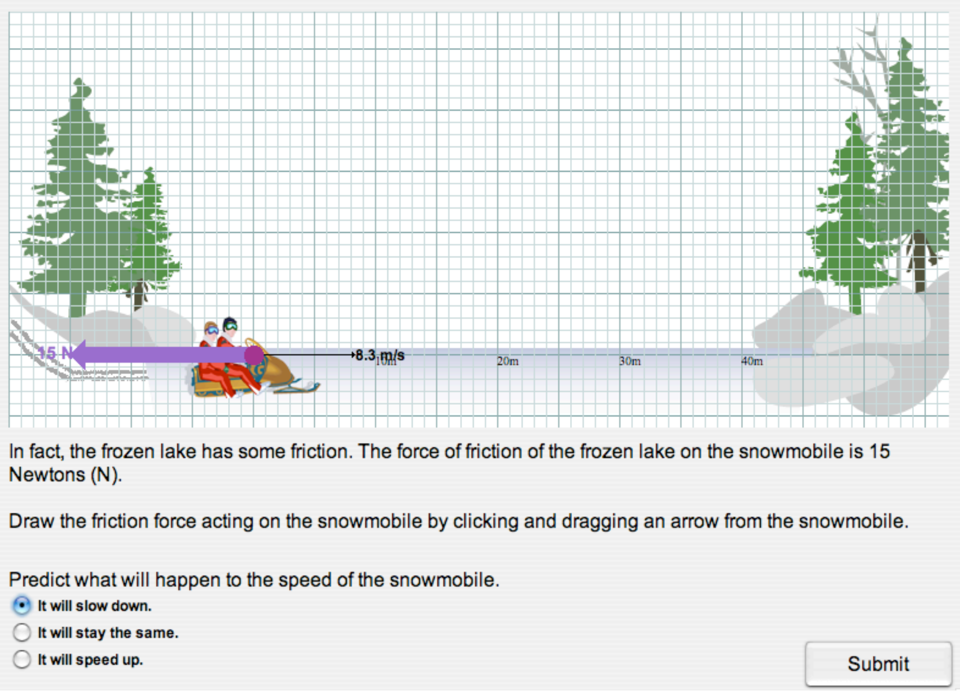
Design principles shaping the Calipers assessment tasks included: (1) specification of a driving, authentic problem, (2) creation of items and tasks to take advantage of the simulation technology, (3) alignment with standards, and (4) alignment with the types of problems and activities presented in curricula.

Simulation-Based Assessments for Forces and Motion

The environment 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 their existing *Dynamica* engine which modeled Newtonian laws of motion. 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 nine physical science. Questions asked students 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.

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Figure 1 presents a screen shot of a scene within one of the Mountain Rescue assessments.



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.

Figure 1. Force and Motion Assessment 1 friction force drawing and prediction items.

Students are asked to draw an arrow (shown in purple) 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.

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 a brief report. The computer records the magnitude and direction of arrows drawn and logs student 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 records the speed selected for each experimental trial. This information can be used to examine how each student and an entire class perform an experiment--a task that cannot be done in a classroom laboratory. Rubrics evaluate whether students have chosen experimental values that cover the range necessary, and if they were systematic in exploring the range of values. Finally, rubics evaluate if students were successful in accomplishing the task.

For many types of responses (i.e., multiple choice, drawing force arrows), the computer can automatically produce a score based on a rule created by the project staff. For example, in

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

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 3 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 injured skiers on an 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, the program averages all the runs that a student makes. In addition, the computer program takes into account whether students identify the target speed and whether they repeat any trials.

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 which the computer captures and integrates into the students’ records.

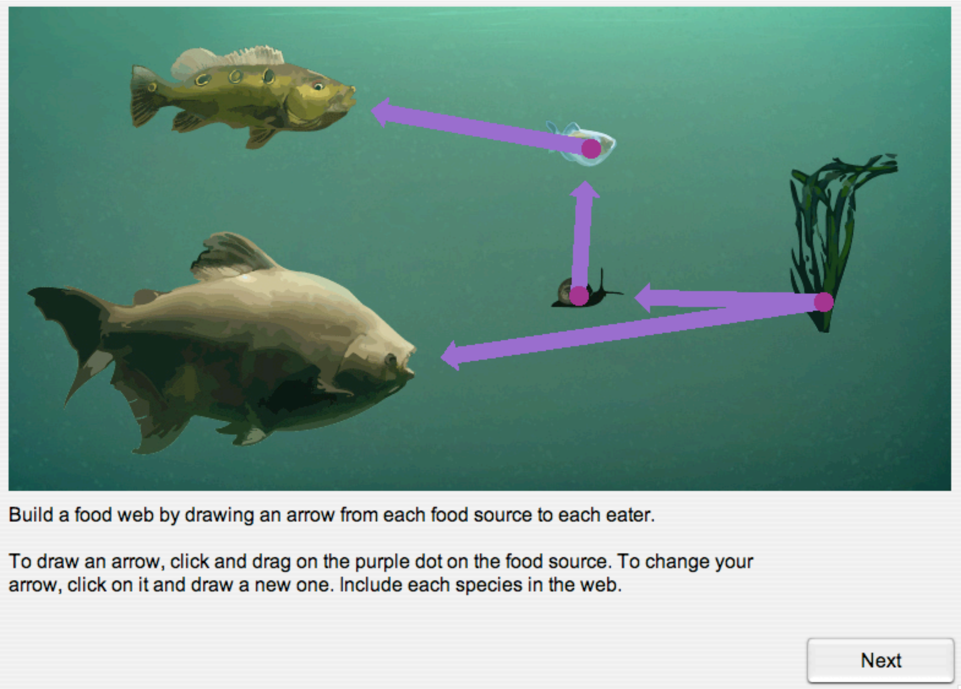
When all of the responses have been scored by computer and humans, the results are placed in a database that can be explored in a variety of ways. A teacher or researcher can see how well students are performing on specific content or inquiry targets or how well students are performing on the assessment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

Simulation-Based Assessments for Ecosystems

The environment 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 their 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 roles and relationships of the fish and plant species and predict and explain the effects of changing the numbers of organisms. Manipulations of the simulation include drawing food webs and varying the number of predator and prey as students run the simulation.

Figure 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 3 presents a screen shot of the population level of the ecosystem in which students vary the numbers of organisms.

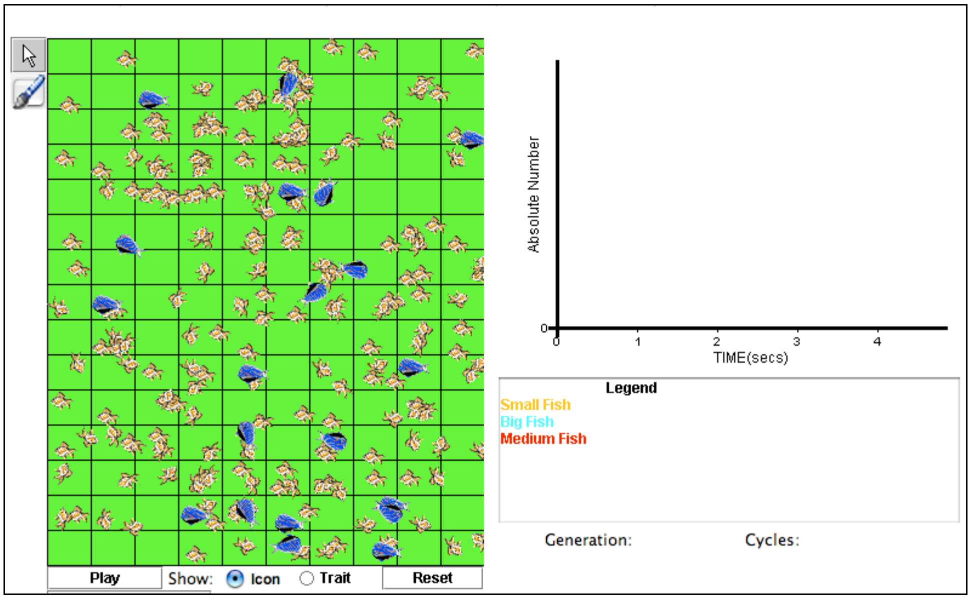
This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Contact: equellm@wested.org



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.

Figure 2. Observing species in the ecosystem.



Legend

Small Fish
Big Fish
Medium Fish

Generation:

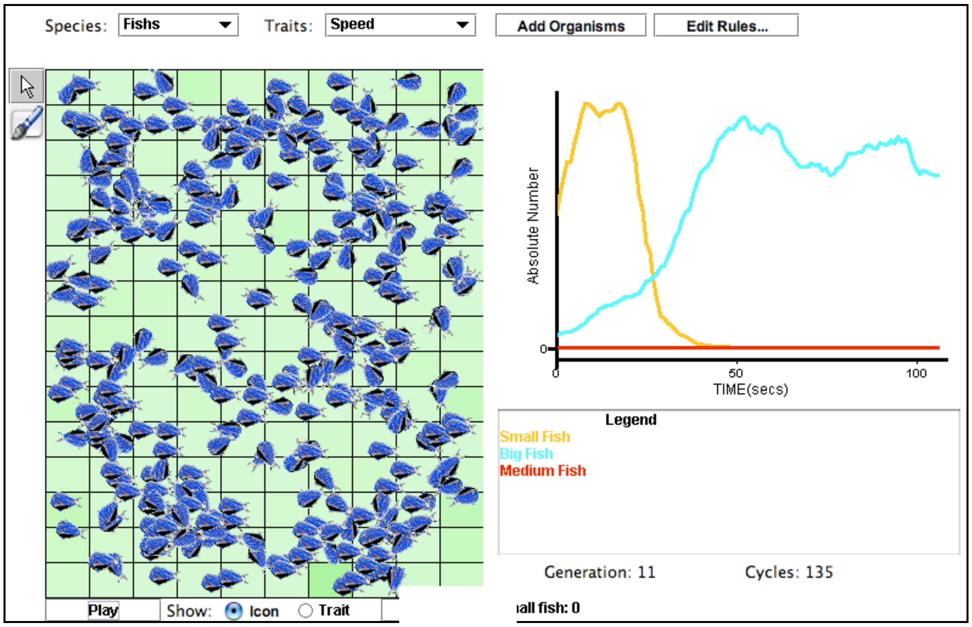
Cycles:

Show: ● Icon ○ Trait

Figure 3. Population view of Fish World

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Students can experiment with different starting populations in order to determine the food web in this newly discovered lake. One example is shown below in Figure 4.



Species: Fishs

Traits: Speed

Show: ● Icon ○ Trait

Generation: 11

Cycles: 135

ıall fish: 0

Figure 4. Population dynamics in Fish World

At the beginning of the simulation there were 100 small fish and 10 big fish. As the simulation runs, the number of both kinds of fish increase, but then the numbers of small fish rapidly decline to zero. The big fish settle into a stable population size for the environment. The questions to students are “Why? What is the food web here? What is your evidence?” This tests students’ ability both to reason from evidence and to conduct experiments that provide the necessary evidence. In other assessment questions, students use sliders to manipulate the number of a predator fish and to predict and explain the effect on other species

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 standard, 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.

Technical Quality Methods

The technical quality evidence being gathered for the Calipers simulation-based assessments includes methods recommended by research and standards for test development: alignment of the assessments with national standards for science, task specifications, expert review of alignment with standards and of content and item quality, analyses of teacher and student data gathered from classroom pilot testing, and cognitive analyses of students thinking-aloud, (AERA/APA/NCME, 2002; Pellegrino, et. al, 2001; Quellmalz, et. al, 2005).

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Alignment of Assessments with Standards and Representative Curriculum Programs. As a first step in establishing content and construct validity, the Calipers project staff aligned the assessment tasks and questions with the AAAS key ideas and the NSES for the targeted content and inquiry abilities for each of the assessments. The assessments were also aligned with four typical middle school science curricula (two conventional textbook-based, two NSF- funded) to confirm the curricular relevance of the assessments. The curriculum analyses desribed the standards, contexts, and types of tasks and questions in the programs. These analyses served as a reference for the design of the Calipers assessment tasks.

Evidence-Centered Design. The Calipers project used evidence-centered assessment design methods to produce re-usable task templates laying out the connections of targeted science knowledge and skills (student model) to features of the simulation environment and assessment questions that would elicit evidence of the skills (task model), and the scores that would calibrate the levels of student knowledge and inquiry skills (evidence model) (Mislevy & Haertel, 2006). Specifications for unique simulation-based, end-of-unit benchmark assessments were prescribed in simulation shells, that in turn informed the design of scenes in storyboards that sketched the layout, functionality, and items to appear on each simulation screen. Then, technology programmers developed the simulation-based prototype assessments for online delivery.

Expert Reviews. A packet of design documents for the Force and Motion assessments (alignment tables, simulation shells, and the actual assessments) was reviewed by AAAS and also by external science experts for quality of the items' science content and inquiry skills and for attention to principles of universal design. Minor revision suggestions are being implemented.

Feasibility Testing. Feasibility testing was conducted for each of the three Force and Motion assessments with at least five middle school students and a high school student. In these feasibility tests, the students participated in think-alouds as they completed the assessments. Findings from the feasibility testing showed that students finished the assessments in the allotted time, used the intended concepts and inquiry skills, and found them engaging. Only minor revisions were required.

Pilot Tests. Pilot tests were conducted by three middle school teachers in multiple classes for the two assessments targeting the middle school Force and Motion standards. Data gathered included teacher interviews, teacher questionnaires on opportunity to learn, teacher classification of students into levels of science achievement, and student responses to the assessments, and cognitive labs.

Results and Findings. Preliminary results and findings are summarized below.

Teacher Interviews. Teachers were interviewed about their perceptions of the Calipers Mountain Rescue assessments and simulation-based assessments, in general. In response to the question, "How well do the Calipers simulation-based assessments tell you about what your students know about science content and inquiry?," Teacher 2 remarked that this scenario-based assessment is testing whether her students "really get it" as opposed to just "memorizing the concepts." Similarly, Teacher 3 said that the Calipers assessment "tells me who knows things at the surface level and who can really think past just this is the equation and then think past how do I use this equation and why is it important in a different situation." Teacher 4 acknowledged that his students had not been exposed to some of the science content addressed in the Calipers

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

assessments but believed that the Calipers assessments could be useful for not only assessing his students, but also help them to learn the content.

Generally, teachers believed that simulation-based assessments are “difficult…but a good challenge,” “interesting,” and “compelling” for students. Teacher 4 stated, “I think my students are extremely engaged in the [Calipers] activity.” Teachers believed that simulation-based assessments are engaging for students because they provide the opportunity for students to use their science knowledge in a “real life” context and they make science meaningful. According to Teacher 2, “[Simulation-based assessments] answer the ‘Why do we care? Why are we learning this?’ questions.”

Opportunity to Learn Questionnaires. All of the teachers addressed most of the science content in the Calipers assessments during partial or full class periods. Content targets related to curved paths were not addressed by two of the teachers, and the other teachers only addressed these targets for less than one class period. All teachers addressed the targets related to most of the science inquiry abilities in their science classes. Tables 1 and 2 show the average science emphasis in terms of instructional time across targets within each content key idea or inquiry ability.

Table 1

Science Content Emphasis

Science Content Key Ideas	**Teacher 1**	**Teacher 2**	**Teacher 3**	**Teacher 4**
Distance, speed, and acceleration	3 (3)	4 (4)	3.6 (4)	4 (4)
Balanced forces	3 (3)	3 (3)	2.5 (2,3)	2.5 (2,3)
Unbalanced forces	3 (3)	2 (2)	2.3 (2)	2.5 (3)
Friction	2 (2)	2 (1,2,3)	2.3 (3)	3 (4)
Curved paths	1 (1)	1 (1)	1.5 (1)	2 (2)

Notes. Teachers responded to 5-point survey items about the number of class periods spent on targets within each of the five forces and motion key ideas (1 = “0 classes”, 2 = “<1 class”, 3 = “1-2 classes”, 4 = “3-4 classes”, and 5 = “>4 classes”). The modal response across all targets within a key idea is reported in parentheses, when applicable. *Distance, speed and acceleration* has 5 targets. *Balanced forces* has 2 targets. *Unbalanced forces* has 6 targets. *Friction* has 3 targets. *Curved paths* has 4 targets.

Table 2

Science Inquiry Emphasis

Science Inquiry Abilities	**Teacher 1**	**Teacher 2**	**Teacher 3**	**Teacher 4**
Identify questions that can be answered through scientific investigations	2.3 (2)	5 (5)	2.3 (3)	5 (5)
Design and conduct a scientific investigation	3.5 (3,4)	3.5 (2,5)	3.3 (3)	2.3 (2)
Use appropriate tools and techniques to gather, analyze, and interpret data	3 (3)	2 (2)	5 (5)	3 (3)
Develop descriptions, explanations, and predictions and models using evidence	4 (4)	4 (5)	5 (5)	3 (3)
Think critically and logically to make the relationships between evidence and	3 (3)	3.5 (2,5)	4 (4)	2 (2)

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

explanations				
Recognize and analyze alternative explanations and predictions	Missing	2 (2)	3 (3)	3 (3)
Communicate scientific procedures and explanations	3 (3)	2 (2)	4 (4)	3 (3)
Use mathematics in all aspects of scientific inquiry	3 (3)	4 (5)	4.3 (5)	3.7 (3)

Notes. Teachers responded to 5-point survey items about the number of class periods spent on targets within each of the eight science inquiry key ideas (1 = “0 classes”, 2 = “<1 class”, 3 = “1-2 classes”, 4 = “3-4 classes”, and 5 = “>4 classes”). The modal response across all targets within a key idea is reported in parentheses, when applicable. *Identify questions that can be answered through scientific investigations* has 3 targets. *Design and conduct a scientific investigation* has 4 targets. *Use appropriate tools and techniques to gather, analyze, and interpret data* has 2 targets. *Develop descriptions, explanations, and predictions and models using evidence* has 4 targets. *Think critically and logically to make the relationships between evidence and explanations* has 4 targets. *Recognize and analyze alternative explanations and predictions* has 2 targets. *Communicate scientific procedures and explanations* has 2 targets. *Use mathematics in all aspects of scientific inquiry* has 3 targets.

Instructional Approaches**.** The most frequently used instructional approaches by teachers were hands-on/laboratory activities and small group work. Table 3 shows the type and frequency of instructional approaches used by each teacher during the Forces and Motion unit.

Table 3

Type and Frequency of Teachers’ Instructional Approaches in Forces and Motion Unit

	Teacher 1	**Teacher 2**	**Teacher 3**	**Teacher 4**
Hands-on/laboratory activities	1-3 times/wk	1-3 times/wk	1-3 times/wk	Almost everyday
Projects that take a week or more	1-3 times/mo	Sometimes	1-3 times/mo	Sometimes
Write in a journal	Never	Never	Never	Never
Suggest or help plan classroom activities	Never	Sometimes	Sometimes	Never
Work in small groups to come up with a joint solution or approach to a problem or task	1-3 times/wk	Sometimes	1-3 times/wk	Almost everyday
Work on problems for which there is no obvious method of solution	Sometimes	Sometimes	Sometimes	Missing
Write an essay in which to explain their thinking or reasoning	Sometimes	Sometimes	Sometimes	1-3 times/wk
Use computers to read information	1-3 times/mo	Never	Never	1-3 times/mo
Use computers to watch videos	Never	Sometimes	Sometimes	Sometimes
Use computers for interactive activities	Sometimes	Sometimes	Never	1-3 times/mo

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Use computers to take multiple-choice assessments	Never	1-3 times/wk	Never	Sometimes
Use computers to take simulation-based assessments (not including Calipers simulations)	Never	Never	Never	Missing

Teachers also reported on the extent to which technology was integrated into their Forces and Motion unit. Technology was infrequently integrated into the Forces and Motion unit for Teacher 1's classes, moderately integrated into classes for Teachers 3 and 4, and frequently integrated into Teacher 2's classes.

Rater training to Score Student Assessment Data. Raters participated in training sessions prior to scoring each item. For each item, raters first discussed the rubric and scored approximately 5 papers together. Raters then completed approximately 6-7 "qualifying" papers. Each qualifying paper was scored by scorers individually. After scoring each item, raters discussed the scores and resolved any discrepancies. Raters then double-scored approximately 30% of papers for each item. The remaining papers were single-coded. Interrater agreement was 80% or higher for most items and all discrepancies were resolved via consensus or by a third rater who also participated in the training session for the item. Items for which interrater agreement was lower than 80% will need revisions to the item prompt or rubric.

Student Performance. *Mountain Rescue 1* addressed four science content key ideas and included multiple-choice, constructed-response and technology-based item formats (e.g., drawing arrows, using sliders). *Mountain Rescue 2* addressed five content key ideas and also included multiple-choice, constructed-response and technology-based item formats. We note that only a portion of the technology-based moves by the student were scored. Tables 4 and 5 show the distribution of items by format and science content key idea for *Mountain Rescue 1* and *Mountain Rescue 2*.

Table 4

Number of Items by Format and Science Content Key Idea for Mountain Rescue 1

Item Format	**Science Content Key Ideas**				
	1	**2**	**3**	**4**	**5**
Multiple Choice	9	1	0	3	1
Constructed Response	12	1	0	4	6
Technology Based	0	0	0	2	1
Total	21	2	0	9	8

Note. Key Idea 1 = Distance, speed, and acceleration . Key Idea 2 = Balanced forces. Key Idea 3 = Unbalanced forces. Key Idea 4 = Friction. Key Idea 5 = Curved paths. One multiple-choice item and one constructed-response item were double-coded as Balanced forces and Friction.

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Table 5

Number of Items by Format and Science Content Key Idea for Mountain Rescue 2

Item Format	**Science Content Key Ideas**				
	1	**2**	**3**	**4**	**5**
Multiple Choice	10	1	2	0	1
Constructed Response	10	0	6	2	5
Technology Based	0	0	1	1	1
Total	20	1	9	3	7

Note. Key Idea 1 = Distance, speed, and acceleration . Key Idea 2 = Balanced forces. Key Idea 3 = Unbalanced forces. Key Idea 4 = Friction. Key Idea 5 = Curved paths. One multiple-choice item and one constructed-response item were double-coded as Balanced forces and Friction.

Mountain Rescue 1 and *Mountain Rescue 2* both addressed five science inquiry key ideas Tables 6 and 7 show the distribution of items by format and science inquiry ability for the assessments.

Table 6

Number of Items by Format and Science Inquiry Ability for Mountain Rescue 1

Item Format	**Science Inquiry Key Abilities**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Multiple Choice	1	2	0	4	0	0	0	4
Constructed Response	1	2	0	4	0	0	9	2
Technology Based	0	1	0	2	0	0	0	0
Total	2	5	0	10	0	0	9	8

Key Ability 1 = Identify questions that can be answered through scientific investigations. Key Ability 2 = Design and conduct a scientific investigation. Key Ability 3 = Use appropriate tools and techniques to gather, analyze, and interpret data. Key Ability 4 = Develop descriptions, explanations, and predictions and models using evidence. Key Ability 5 = Think critically and logically to make the relationships between evidence and explanations. Key Ability 6 = Recognize and analyze alternative explanations and predictions. Key Ability 7 = Communicate scientific procedures and explanations. Key Ability 8 = Use mathematics in all aspects of scientific inquiry.

Table 7

Number of Items by Format and Science Inquiry Ability for Mountain Rescue 2

Item Format	**Science Inquiry Key Abilities**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Multiple Choice	1	2	0	0	0	0	0	8
Constructed Response	1	2	0	4	0	0	9	0
Technology Based	0	1	0	2	0	0	0	0
Total	2	5	0	6	0	0	9	8

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Key Ability 1 = Identify questions that can be answered through scientific investigations. Key Ability 2 = Design and conduct a scientific investigation. Key Ability 3 = Use appropriate tools and techniques to gather, analyze, and interpret data. Key Ability 4 = Develop descriptions, explanations, and predictions and models using evidence. Key Ability 5 = Think critically and logically to make the relationships between evidence and explanations. Key Ability 6 = Recognize and analyze alternative explanations and predictions. Key Ability 7 = Communicate scientific procedures and explanations. Key Ability 8 = Use mathematics in all aspects of scientific inquiry.

Tables 8 and 9 show average student performance overall and by science content key idea for high, medium, and low science achievers (based on teacher ratings) for *Mountain Rescue 1* and *Mountain Rescue 2*. All constructed response items were recoded on a scale from 0-1 for the purpose of these analyses. Since students had so little opportunity to learn the curved path concepts, items related to curved path key ideas are not included in the analyses of student performance. In addition, key ideas that were not addressed in the assessment are not included in the tables. For most of the content areas, high science achievers performed best and low science achievers performed least well. On *Mountain Rescue 1*, these differences in performance based on prior achievement appear most prominent for the Balanced Forces and Friction key ideas and this pattern was not observed for the Distance, Speed, Acceleration key idea.

Table 8

Student Performance on Science Content Key Ideas for Mountain Rescue

		Science Content Key Ideas		
		1	**2**	**4**
Overall	**n**	109	100	109
	M (SD)	.57 (.26)	.50 (.34)	.55 (.26)
High Achievers	**n**	37	34	37
	M (SD)	.67 (.26)	.57 (.33)	.64 (.23)
Med Achievers	**n**	46	41	46
	M (SD)	.52 (.25)	.53 (.31)	.52 (.28)
Low Achievers	**n**	25	24	25
	M (SD)	.52 (.22)	.32 (.36)	.46 (.23)

Note. Key Idea 1 = Distance, speed, and acceleration . Key Idea 2 = Balanced forces. Key Idea 4 = Friction.

Table 9

Student Performance on Science Content Key Ideas for Mountain Rescue 2

		Science Content Key Ideas		
		1	**3**	**4**
Overall	**n**	108	108	99
	M (SD)	.55 (.25)	.20 (.17)	.35 (.23)
High Achievers	**n**	31	31	29
	M (SD)	.67 (.22)	.23 (.16)	.45 (.17)
Med Achievers	**n**	35	35	32
	M (SD)	.54 (.25)	.19 (.18)	.35 (.18)
Low Achievers	**n**	27	27	24
	M (SD)	.44 (.25)	.14 (.13)	.22 (.27)

Note. Key Idea 1 = Distance, speed, and acceleration . Key Idea 3 = Unbalanced forces. Key Idea 4 = Friction.

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Tables 10 and 11 show average student performance overall and by science inquiry ability for high, medium, and low science achievers (based on teacher ratings) for *Mountain Rescue 1* and *Mountain Rescue 2*. All constructed response items were recoded on a scale from 0-1 for the purpose of these analyses. Since students had so little opportunity to learn the curved path concepts, items related to curved path key ideas are not included in the analyses of student performance. In addition, findings on the multiple choice item related to key idea 2 are not reported because of significant missing data for this item. Again, high science achievers receive the highest scores, and low science achievers receive the lowest scores on items related to each of the inquiry abilities on average. These findings provide evidence of the discriminant validity of the items.

Table 10
Student Performance on Science Inquiry Abilities for Mountain Rescue 1

		Science Inquiry Abilities				
		1	**2**	**4**	**7**	**8**
Overall	**n**	94	109	109	102	109
	M (SD)	.48 (.40)	.46 (.25)	.58 (.27)	.52 (.26)	.57 (.28)
High Achievers	**n**	31	37	37	34	37
	M (SD)	.66 (.33)	.50 (.25)	.68 (.25)	.62 (.24)	.71 (.26)
Med Achievers	**n**	40	46	46	43	46
	M (SD)	.44 (.41)	.44 (.24)	.55 (.30)	.52 (.24)	.52 (.26)
Low Achievers	**n**	23	25	25	24	25
	M (SD)	.35 (.42)	.43 (.25)	.51 (.24)	.41 (.27)	.49 (.27)

Note. Key Ability 1 = Identify questions that can be answered through scientific investigations. Key Ability 2 = Design and conduct a scientific investigation. Key Ability 4 = Develop descriptions, explanations, and predictions and models using evidence. Key Ability 7 = Communicate scientific procedures and explanations. Key Ability 8 = Use mathematics in all aspects of scientific inquiry.

Table 11
Student Performance on Science Inquiry Abilities for Mountain Rescue 2

		Science Inquiry Key Abilities				
		1	**2**	**4**	**7**	**8**
Overall	**n**	99	108	108	102	108
	M (SD)	.57 (.37)	.48 (.26)	.29 (.20)	.39 (.23)	.55 (.24)
High Achievers	**n**	29	31	31	30	31
	M (SD)	.67 (.40)	.58 (.21)	.35 (.17)	.47 (.20)	.65 (.22)
Med Achievers	**n**	32	35	35	33	35
	M (SD)	.63 (.31)	.46 (.27)	.29 (.19)	.42 (.22)	.53 (.25)
Low Achievers	**n**	24	27	27	25	27
	M (SD)	.48 (.35)	.37 (.27)	.18 (.19)	.31 (.23)	.43 (.20)

Note. Key Ability 1 = Identify questions that can be answered through scientific investigations. Key Ability 2 = Design and conduct a scientific investigation. Key Ability 4 = Develop descriptions, explanations, and predictions and models using

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

evidence. Key Ability 7 = Communicate scientific procedures and explanations. Key Ability 8 = Use mathematics in all aspects of scientific inquiry.

Tables 12 and 13 show average student performance overall and by item format for *Mountain Rescue 1* and *Mountain Rescue 2*. The three technology-based items are the most challenging in the assessments.

Table 12
Student Performance by Item Format for Mountain Rescue 1

		Item Format		
		Multiple Choice (n = 14)	**Constructed Response (n = 24)**	**Technology Based (n = 3)**
Overall	**n**	109	105	109
	M (SD)	.66 (.24)	.48 (.24)	.40 (.26)
High Achievers	**n**	37	35	37
	M (SD)	.75 (.24)	.62 (.22)	.47 (.23)
Med Achievers	**n**	46	44	46
	M (SD)	.61 (.25)	.44 (.21)	.37 (.27)
Low Achievers	**n**	25	25	25
	M (SD)	.62 (.19)	.39 (.23)	.39 (.25)

Table 13
Student Performance by Item Format for Mountain Rescue 2

		Item Format		
		Multiple Choice (n = 14)	**Constructed Response (n = 21)**	**Technology Based (n = 3)**
Overall	**n**	108	104	108
	M (SD)	.58 (.24)	.42 (.23)	.29 (.21)
High Achievers	**n**	31	30	31
	M (SD)	.71 (.20)	.54 (.20)	.32 (.20)
Med Achievers	**n**	35	34	35
	M (SD)	.57 (.23)	.42 (.20)	.29 (.21)
Low Achievers	**n**	27	26	27
	M (SD)	.45 (.22)	.34 (.27)	.18 (.15)

Analyses of the student assessment data and cognitive labs are currently in progress. The cognitive analyses will be used to contribute documentation of the construct and content validity of the assessments.

The development of the Ecosystem assessments is employing technical quality procedures parallel to those used for the Force and Motion assessments. The Ecosystem assessments have been aligned with standards and curricula. They will be submitted for expert review. They are scheduled for pilot testing in spring, 2007.

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

Additional Analyses. Basic technical quality information will be reported for each item including the number of students tested, the range and distribution of scores for each question, the number of individuals who performed at each level of the assessment rubric, p-values, and the standard deviation and standard error of measurement. Quantitative analyses to investigate the relationships between opportunity to learn and student performance will be conducted. Correlations of exposure to the item content and performance on the item will be calculated and provide evidence of the instructional sensitivity of the assessment items.

The Promise of Simulation-Based Science Assessments

The Calipers demonstration project aims to provide evidence of the feasibility, technical quality, and utility of simulation-based science assessments. Development of environments modeling scientifically-based principles will allow the use and re-use of the underlying programming of the environment for both assessment and instruction. For example, the ecosystem environment can be adapted for other aquatic (e.g., salt water) 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 importantly, 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.

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, NCME). (2002). *Standards for educational and psychological testing*. Washington, DC: Author.

Baxter, G. P. (1995). Using computer simulations to assess hands-on science learning. *Journal of Science Education and Technology, 4*, 21-27.

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

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (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., 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*, 495-538.

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

Jackson, S., Stratford, S., Krajcik, J., & Soloway, E. (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., 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.

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.

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

Quellmalz, E. S., Haertel, G. D., DeBarger, A. H., & Kreikemeier, P. (2005). *A study of evidence of the validities of assessments of science inquiry in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the New Standards Science Reference Exam (NSSRE) in Science* (Validities Technical Report #1). Menlo Park, CA: SRI International.

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.

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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

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

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

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

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

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

CONTACT: Edys Quellmalz
WestEd
400 Seaport Ct. Suite 222
Redwood City, CA 94063
equellm@wested.org

This material is based upon work supported by the National Science Foundation under Grant No. 0454772 awarded to SRI International, Edys Quellmalz, Principal Investigator. 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