The promise of simulation-based science assessment: the Calipers project

Edys S. Quellmalz*, Michael J. Timms and Barbara Buckley

Math, Science and Technology Program,
WestEd, 400 Seaport Court, Suite 222,
Redwood City, CA 94063, USA
E-mail: equellm@wested.org
E-mail: mtimms@wested.org
E-mail: bbuckle@wested.org
*Corresponding author

Abstract: The Calipers project developed and studied a new generation of simulation-based assessment systems. The project aimed to demonstrate the potential of technology- and simulation-based assessments to provide high-quality evidence of complex performances for science tests that address accountability or formative goals. End-of-unit, benchmark assessments for the topics of ecosystems and for forces and motion were developed to test national science standards at the middle school and secondary levels. Technical quality evidence documented the alignment of the assessments with national science standards, expert reviews of content and item quality, cognitive analyses of students thinking-aloud, and analyses of teacher and student data gathered from classroom pilot testing. The project broke new ground in harnessing the affordances of technology to transform what, how, when and where science learning is assessed and to gather evidence of students' connected science knowledge and extended inquiry not well measured by traditional paper-based tests.

Keywords: assessment; science; technology; simulations.

Reference to this paper should be made as follows: Quellmalz, E.S., Timms, M.J. and Buckley, B. (2010) 'The promise of simulation-based science assessment: the Calipers project', *Int. J. Learning Technology*, Vol. 5, No. 3, pp.243–263.

Biographical notes: Edys Quellmalz is the Director of WestEd's Technology Enhanced Assessments and Learning Systems. She directs research, development and evaluation projects funded by the National Science Foundation (NSF): 'Calipers I: Using simulations to assess complex science learning', 'Calipers II: Using simulations to assess complex learning, Foundations of 21st century science assessments, Evaluation of the Bioinformatics project'; and the US Department of Education Institute of Educational Science (IES): 'Multilevel assessments of science standards, and SimScientists'.

Mike Timms is the Associate Director of WestEd's Mathematics, Science and Technology Program. He is Co-Principal Investigator on projects funded by the National Science Foundation (NSF): 'Calipers II: Foundations of 21st century science assessments, Evaluation of Bioinformatics'; and the US Department of Education Institute of Educational Science (IES): 'Multilevel assessments of

science standards, SimScientists' and 'Evalution of the FOSS ASK project'. He has managed large scale assessment development projects and conducted evaluations of programs that have used technology as a component in enhancing teacher, student and museum visitor understanding. He specialises in assessment development, evaluation of educational technology, computer-based development, educational administration and project management.

Barbara Buckley is a Senior Research Associate at WestEd, where she leads content teams in biological, earth and physical science as they develop simulation-based assessments and inquiry activities for middle school classrooms. Her primary focus is on the use of technology for supporting model-based learning and assessment in science.

1 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 that can be tested, how, where and when testing can take place, and the evidence available for monitoring student learning (Quellmalz and 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. Simulations can be used to test students' knowledge of science system components and interrelationships. Because simulations are dynamic and can be manipulated, students can demonstrate their abilities to engage in active inquiry. Technology-based assessments can elicit, collect, document, analyse, appraise, support and display kinds of student performances that have not been readily accessible through traditional paper-based testing methods. As online testing increases, computer simulations are being seen as a potentially more affordable option for formative and summative assessments that can be designed for both classroom and large-scale administration. Simulations can offer several advantages over hands-on tasks in terms of costs (Baxter, 1995), and they can be administered to a large sample of students simultaneously or to individuals or small groups just-in-time. Computer systems can easily capture student responses and produce a variety of standards-based performance reports at the individual, class, school, district, and state levels. By incorporating simulation-based assessments, science assessment systems can add assessment data that are not available from conventional test formats.

Currently, most technology-based accountability assessments do not incorporate complex performance tasks, nor do most technology-rich curricula yet employ principled assessment designs that provide student performance data on how well students can use scientific knowledge to conduct scientific inquiry. In this paper, we describe a project funded by the National Science Foundation.

This material is based upon work supported by the National Science Foundation under Grant No. ESI-0454772 awarded to SRI International and No. 0733345 awarded to WestEd, 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.

Foundation, ‘Calipers: Using simulations to assess complex science learning’, which has developed assessment designs and prototypes that can take advantage of technology to bring assessments of complex performances that meet the testing standards for technical quality required for sound assessment practice into science tests with either accountability or formative goals.

2 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’ and actively investigate a variety of concepts and relationships through interactions with multiple representations (e.g., pictures, models, graphs, tables). These affordances of simulations may permit students from diverse language backgrounds and learning styles to better understand the demands of assessment tasks and questions and to provide students with 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 of spatial, causal and temporal relationships. Technology allows students to manipulate an array of variables, observe their impacts and try again.

Simulations can allow students to engage in the kinds of investigations that are familiar components of hands-on curricula and also to explore realistic problem scenarios that are difficult or impossible to create in typical classrooms. Simulations allow experimentation with phenomena that are too large or small, fast or slow, or too expensive or dangerous. In addition, use of simulations can overcome many of the economic and logistical constraints associated with the purchase, replenishment and setting up of equipment for hands-on science experiments.

3 Research on simulations and student learning

Numerous studies have discussed the benefits of using simulations to support student learning. Model-It was used in a large number of classrooms, and positive learning outcomes based on pretest-post-test 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 them (Jackson et al., 1995). 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 and Wilensky, 2003). Middle school students who completed the ThinkerTools curriculum performed better on average on basic physics problems than high school students 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 showed an increase in genetics content knowledge in specific areas, as well as an increase in genetics problem-solving skills (Buckley, 2004). The Modelling Across the Curriculum (MAC) project expanded this research into a large-scale study of high school students' model-based learning with computer models. Project data from 2005–2006 showed significant learning gains in a majority of biology, physical science, physics and chemistry classes. In addition, the MAC project permitted analyses of students' problem solving strategies collected by log file data. In-process log data revealed if students successfully completed tasks and also allowed diagnoses at what steps students had difficulties and whether students solved problems systematically (Horwitz et al., 2007; see also Buckley, this issue).

4 The Calipers project goals

The Calipers project was a two-year demonstration project that used evidence-centred assessment design methods to develop technology-supported 'benchmark assessments' with technical quality to bridge the gap between external, summative assessments and curriculum-embedded, formative assessments (Mislevy and Haertel, 2006). The Calipers project developed a new generation of technology-based science assessments to 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 were 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 demonstration project documented the feasibility, usability and technical quality of the new simulation-based assessments. Feasibility and usability of the assessments were established first through testing the assessments with a small number of students. Feasibility tests confirmed that students could complete the assessments in the allotted times and that they were addressing the intended science content. Usability testing established that students could navigate through the simulation-based assessments and respond to questions in the modules. The technical quality of the assessments was established by gathering evidence from pilot tests, opportunity to learn surveys, teacher surveys and interviews, expert reviews of alignments with national science standards and item quality and analysis of student responses to the items in the assessment gathered during cognitive labs and classroom pilot testing. An import aspect of the Calipers project was that it also examined the use of non-scored variables, such as student time spent on an activity or number of trials run, and showed that such variables are worth measuring in addition to the traditionally scored proficiency variables.

5 Development of the Calipers assessments

The Calipers assessments were shaped by a rigorous approach to the assessment design, aligned with key national science standards and representative science curricula, pilot tested and revised. The Calipers assessments were linked to key strands in the AAAS

Atlas of Science Literacy and core National Science Education Standards (NSES) in life science related to populations and ecosystems and in physical science related to forces and motion. Design of the assessments followed an evidence-centred design framework that linked the knowledge and skills to be tested (student model), to features of tasks in which students could 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 alignment with national science standards

2 specification of a driving, authentic problem

3 creation of items and tasks that would take advantage of the simulation technology

4 alignment with the types of problems and activities presented in widely used curricula.

5.1 Assessment item types

The simulation-based assessments included a mixture of *selected response*, *constructed response* and *technology-based* item formats. Table 1 shows the average percentage of the three item types in each assessment. *Selected response* items included any items where the student selected an answer from a set of choices. In the force and motion assessment, just over one third of the items used to assess science content and science inquiry key ideas were selected response. However, selected response items formed a smaller part of the ecosystems assessments, with 13% of the constructed responses testing content and 3% testing inquiry skills. Selected response items were scored automatically by the computer system. *Constructed-response* items were those in which the students made a free response in a text box. Rubrics were developed to score each item. These items formed the bulk of the assessments. In the force and motion assessment, just over half of the items were constructed response, and in the ecosystems assessment, 83% were used to assess content and 69% to assess inquiry. *Technology-based* items included such responses as setting values of input variables in simulations and drawing arrows to represent physical forces or links in a food web. These items were automatically scored according to rules developed for each task.

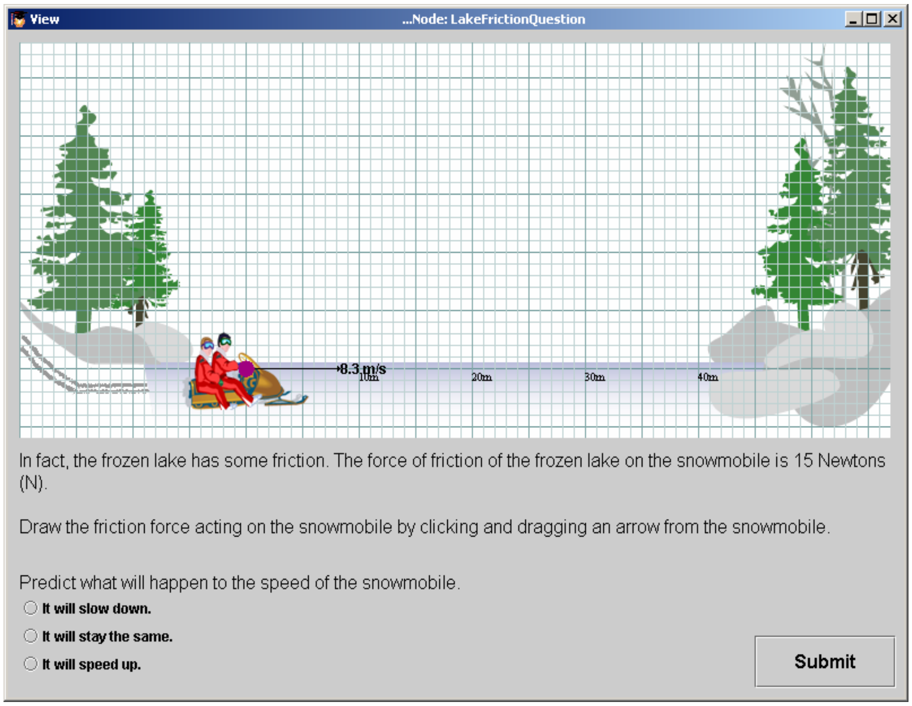
Table 1 Average percentage of items by type used to assess science content and inquiry key ideas

	Science content		*Science inquiry*	
	Force and motion	*Ecosystems*	*Force and motion*	*Ecosystems*
Selected response	35%	13%	36%	3%
Constructed response	57%	83%	54%	69%
Technology-based	8%	4%	10%	28%

5.2 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 title chosen for the simulation-bases assessment module was *Mountain Rescue*, and the driving problem was the need for a student dispatcher to coordinate the rescue of injured skiers by snowmobile units. The simulation environment was developed by Concord Consortium and was built on their existing *Dynamica* engine that modelled Newtonian laws of motion (Horwitz et al., 2007). To demonstrate the flexibility of the environment for assessments at a range of levels of complexity, three assessment tasks 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. The middle school assessment, *Mountain Rescue 1* addressed four science content key ideas, which included distance, speed and acceleration; balanced forces; friction; and curved paths. A high school version of the assessment, *Mountain Rescue 2*, addressed the same four content key ideas and one other, unbalanced forces.

Figure 1 Force and motion assessment 1 – friction force drawing and prediction items (see online version for colours)



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

As students participated in the force and motion assessments, the computer captured their answers to questions in the forms of selected response, short answer, or a brief written report. The computer recorded the magnitude and direction of arrows drawn and logged student manipulations of the simulations. When students experimented with the snowmobile speed to determine the best speed for getting to skiers on an icy hill, the computer recorded the speed selected for each experimental trial. This information could be used to examine how each student and an entire class performed an experiment – a task that could not be done in a classroom laboratory. Rubrics were developed to evaluate whether students had chosen experimental values that covered the range necessary, and if they were systematic in exploring the range of values.

For many types of responses (i.e., selected response, drawing force arrows), the computer automatically produced a score based on a rule created by the project staff. For example, in the first force and motion assessment, students were asked to calculate how long it would take to travel a certain distance at a given speed. Students first selected the correct formula for performing this calculation, then entered the values for distance and speed. The computer calculated the answer and students were asked to evaluate the resulting answer. The computer automatically scored student responses using a rubric that awarded two points for selecting the correct formula the first time, one point for selecting it on the second or third try and zero points for failing to select the correct formula within three tries. A similar rule awarded points for entering the correct values into the equation. If students accurately evaluated the computer's calculations, another point was awarded. In contrast to assessments that score only the final answer, this enabled the assessment to pinpoint where students had difficulty.

When students were conducting experiments to determine the best speed for the snowmobile to use to reach the injured skiers on an icy hill, the score was determined by examining if each experimental value entered was closer to or further away from the 'correct' speed. Students received one point for moving closer to the target speed. For the entire task, the program averaged all the runs that a student made. In addition, the computer program took into account whether students identified the target speed and whether they repeated any trials.

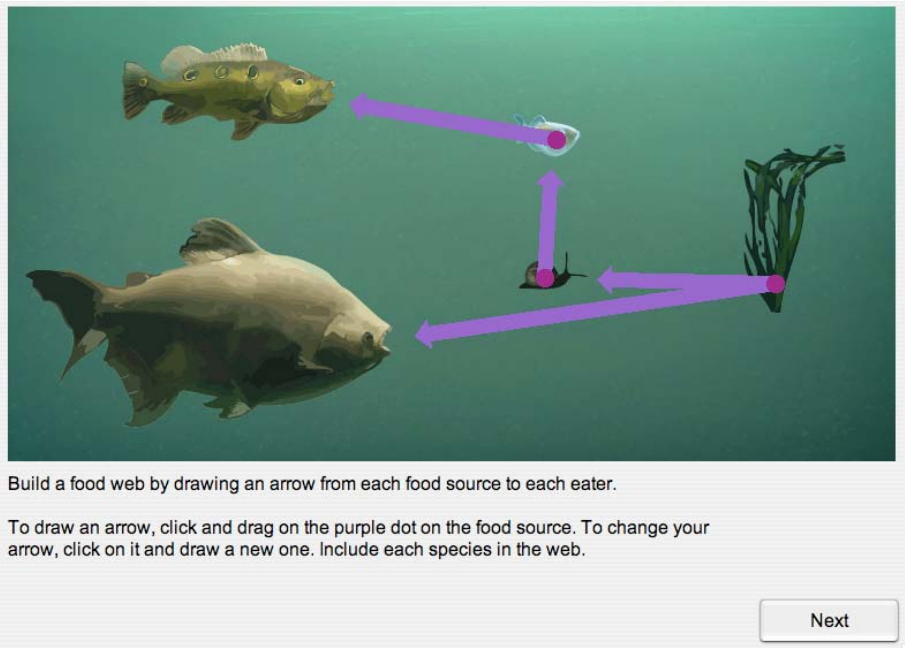
For the constructed-response text-based questions, the computer captured the text exactly as the student typed it. Another program displayed the answers of the entire class, along with the question and the scoring rubric. The researchers read the response, compared it to the rubric, and entered a score, which the computer captured and integrated into the students' records. In a classroom implementation, the teacher would do this scoring.

When all of the responses had been scored by the computer and human raters, the results were placed in a database that could be explored in a variety of ways. A teacher or researcher could see how well students performed on specific content or inquiry targets or how well students performed on the assessment as a whole. Researchers could compare how well students in different curricula performed.

5.3 Simulation-based assessments for ecosystems

The environment selected to simulate principles for populations and ecosystems was a newly discovered lake in the jungle, which was entitled *Fish World* (FW). The driving problem was to explore the lake and describe its ecosystem. The simulation environment for modelling the ecosystem was developed by Concord Consortium extending their existing *Biologica* engine (Horwitz et al., 2007). To demonstrate the flexibility of the simulation environment for assessments at a range of levels of complexity, two assessments were developed to test concepts and inquiry strategies appropriate from the early middle school grades to high school biology. Students were asked to identify the roles and relationships of the fish and plant species and to predict and explain the effects of changing the numbers of organisms. Manipulations of the simulation included drawing food webs and varying the number of predator and prey before students ran the simulation. *FW* addressed five content key ideas, which included diversity of life; food webs; interdependence; adaptation, variation and evolution; and populations.

Figure 2 Ecosystem assessment – drawing food web based on observing species in the ecosystem (see online version for colours)

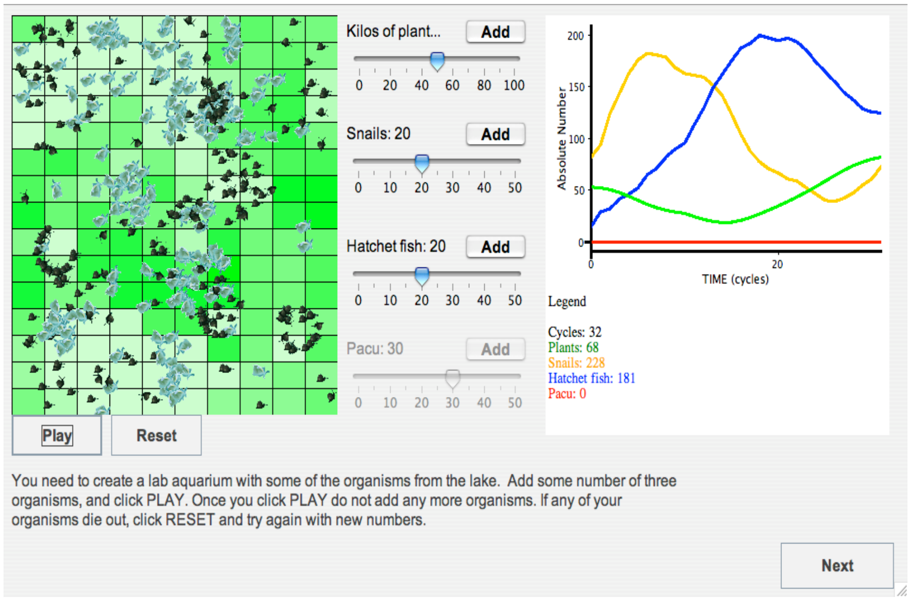


Build a food web by drawing an arrow from each food source to each eater.

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 presents a screen shot of a scene within one of the *FW* assessments, in which students observed unknown species and drew a food web. Figure 3 presents a screen shot of the population level of the ecosystem in which students varied the numbers of organisms. The ecosystem assessments also presented a ‘birds-eye’ population level view of the lake ecosystem. A series of inquiry tasks engaged students in conducting investigations to see how changes in the numbers of the different organisms in the lake affected the ecosystem. One example is shown in Figure 3.

Figure 3 Ecosystem assessment – investigating and interpreting population dynamics in FW (see online version for colours)



Snails: 20

Hatchet fish: 20

Pacu: 30

Legend

Cycles: 32
Plants: 68
Snails: 228
Hatchet fish: 181
Pacu: 0

You need to create a lab aquarium with some of the organisms from the lake. Add some number of three organisms, and click PLAY. Once you click PLAY do not add any more organisms. If any of your organisms die out, click RESET and try again with new numbers.

At the beginning of the simulation there were 50 kilograms of plants, 20 snails and 20 hatchet fish. As the simulation ran, the graph depicted the changing numbers of organisms over time (cycles). In Figure 3, the number of snails increased, followed by an increase in the hatchet fish. Then the number of snails declined as they were eaten by the hatchet fish, which then also declined due to diminished food supply. The plants and snails had an inverse relationship; as the number of snails increased, the amount of plants decreased due to predation. The three organisms may or may not settle into a steady state. The questions to students were, "Describe how the hatchet fish population changed from start to finish. Use evidence from the model," and "Why did this happen?". These questions tested students' ability to interpret patterns represented on the graph, reason from evidence and to conduct experiments that provided the necessary evidence. In other assessment questions, students used 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 students' actions manipulating the simulation were recorded by the computer and scored either automatically or by human raters. The scores could be displayed by concept and inquiry standard, potentially 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 be conducted to document inter-rater reliability data for the constructed response items.

6 Technical quality of the assessments

A goal of the Calipers project was to demonstrate that it is possible to develop simulation-based science assessments that have strong technical qualities (i.e., are both reliable and valid for the purpose of measuring science content knowledge and inquiry skills). This section describes the general sources of data that were collected to provide evidence of this, while Sections 7 and 8 describe in more detail how the data provided evidence of the validity of the assessments. The technical quality evidence gathered for the Calipers simulation-based summative assessments included methods recommended by research and professional 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 et al., 2002; Pellegrino et al., 2001; Quellmalz et al., 2005).

6.1 Pilot tests

For the force and motion assessments, pilot tests were conducted by four middle school teachers in multiple classes for the two assessments targeting the middle school force and motion standards. The two ecosystem assessments, one for middle school, one for secondary biology, were pilot tested in 13 classes taught by three teachers. 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.

6.2 Instructional approaches in the pilot test classrooms

Based on the teachers' responses to our survey of instructional approaches, the most frequently used instructional approaches by the force and motion teachers were hands-on/laboratory activities and small group work. One of the four teachers infrequently integrated technology into the Forces and Motion unit. Two teachers moderately integrated technology into their classes, while another teacher frequently integrated technology into the force and motion unit. The most frequent instructional approaches used by teachers in ecosystems were hands-on/laboratory activities and small group work. Neither teacher indicated that technology was integrated into the ecosystems classes.

6.3 Opportunity-to-learn questionnaires

Teachers also completed opportunity-to-learn questionnaires that provided detailed information about student's exposure to the science knowledge and inquiry skills assessed. Tables 2 and 3 show for the force and motion assessment the average emphasis in terms of instructional time across targets within each content key idea or inquiry ability, which were derived from the AAAS benchmarks and NSES for force and motion. In Tables 2 and 3, MR1 refers to the Mountain Rescue 1 assessment and MR2 to Mountain Rescue 2, which were the force and motion assessments. The four teachers reported addressing most of the science content in the Calipers assessments during partial or full class periods. They devoted the most class time to key ideas related to distance, speed and acceleration, and balanced and unbalanced forces. Friction was generally addressed in one to two 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. As a consequence, we deleted the curved path items from the pilot test analyses. All teachers addressed the targets related to most of the science inquiry abilities in their science classes.

Table 2 Force and motion teachers' science content emphasis compared to number of items in mountain rescue (MR) assessments (n = 4)

Science content key ideas	*Class periods**	*MR1 items*	*MR2 items*
Distance, speed and acceleration	3	21	20
Balanced forces	2	2	1
Unbalanced forces	2		9
Friction	1	9	3
Curved paths	0	8	7

Notes: *Mode of class periods per key idea target, where 0 = no classes; 1 = < 1 class; 2 = 1–2 classes; 3 = 3–4 classes; and 4 = > 4 classes.

Table 3 Force and motion teachers' science inquiry emphasis compared to number of items in mountain rescue (MR) assessments (n = 4)

Science inquiry abilities	*Class periods*	*MR1 items*	*MR2 items*
Identify questions that can be answered through scientific investigations	3	2	2
Design and conduct a scientific investigation	2	5	5
Use appropriate tools and techniques to gather, analyse and interpret data	2	0	0
Develop descriptions, explanations and predictions and models using evidence	3	10	6
Think critically and logically to make the relationships between evidence and explanations	2	0	0
Recognise and analyse alternative explanations and predictions	2	0	0
Communicate scientific procedures and explanations	2	9	9
Use mathematics in all aspects of scientific inquiry	3	8	8

Notes: *Mode of class periods per key idea target, where 0 = no classes; 1 = < 1 class; 2 = 1–2 classes; 3 = 3–4 classes; and 4 = > 4 classes.

For ecosystems, two of the three teachers completed the questionnaire. Tables 4 and 5 show the average science emphasis in terms of instructional time across targets within each ecosystems content key idea or inquiry ability. Both teachers addressed most of the science content in the Calipers assessments during partial or full class periods. It should be noted that Teacher 1 is a high school teacher while Teacher 2 is a middle school teacher. Teacher 1 did not address content targets related to adaptation, variation and evolution and populations, and only addressed targets related to Interdependence for less than one class period. Teacher 2 addressed content targets related to adaptation, variation and evolution for less than one class period. Teacher 1 addressed the targets related to all of the science inquiry abilities in their science classes. Teacher 2 did not address or spent less than one class period addressing most of these targets.

Table 4 Ecosystems teachers' science content emphasis compared to number of items in FW assessments (n = 2)

Science content key ideas	*Class periods* (middle school)*	*FW1 items*	*Class periods* (high school)*	*FW2 items*
Diversity of life	3	1	2	3
Food webs	2	9	2	6
Interdependence	1	0	0	14
Adaptation, variation and evolution	0	0	0	3
Populations	2	12	0	5

Notes: *Mode of class periods per key idea target, where 0 = no classes; 1 = < 1 class; 2 = 1–2 classes; 3 = 3–4 classes; and 4 = > 4 classes.

Table 5 Ecosystem teachers' science inquiry emphasis compared to number of items in FW assessments (n = 2)

Science inquiry abilities	*Class periods* (middle school)*	*FW1 items*	*Class periods* (high school)*	*FW2 items*
Identify questions that can be answered through scientific investigations	1	0	2	0
Design and conduct a scientific investigation	0	1	5	7
Use appropriate tools and techniques to gather, analyse, and interpret data	1	5	5	4
Develop descriptions, explanations, and predictions and models using evidence	0	18	2	19
Think critically and logically to make the relationships between evidence and explanations	1	4	2	1
Recognise and analyse alternative explanations and predictions	0	0	2	1
Communicate scientific procedures and explanations	1	2	3	4
Use mathematics in all aspects of scientific inquiry	1	0	3	0

Notes: *Mode of class periods per key idea target, where 0 = no classes; 1 = < 1 class; 2 = 1–2 classes; 3 = 3–4 classes; and 4 = > 4 classes.

7 Evidence of the validity of the assessments

The evidence of the validity of the assessments presented in this section follows the format for sources of validity evidence described in the Standards for Educational and Psychological Testing (AERA et al., 2002), which include evidence based on test content, response processes, internal test structure and relations to other variables.

7.1 Test content evidence

Technical quality of the test content was established by data from review of the alignment of the items with standards and curricula, expert review of science and item quality and teacher interviews.

7.1.1 Alignment to standards and curricula

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. In addition, the force and motion assessments were aligned with four typical middle school science curricula (two conventional textbook-based, two NSF- funded) to confirm the curricular relevance of the assessments. The ecosystem assessments were similarly aligned with middle school and tenth grade biology textbooks and NSF-funded curriculum projects. The curriculum analyses described the standards, contexts and types of tasks and questions in the programs. These analyses served as one reference for the design of the Calipers assessment tasks.

7.1.2 Expert reviews

The assessment design documents (alignment tables, simulation shells and the actual assessments) were reviewed by AAAS and by additional external science experts for quality of the items' science content and inquiry skills and for attention to principles of universal design. These expert reviews confirmed the alignment of the Calipers tasks and items with their intended national science standards and also confirmed the quality of the assessment items and tasks. Only minor revisions were recommended.

7.1.3 Teacher interviews

Teachers were interviewed about their perceptions of the Calipers simulation-based assessments. Teachers indicated that they thought using the assessments would be practical given sufficient access to computers. Importantly, all the pilot teachers were very positive about the Calipers simulation-based assessments. The teachers felt that simulation-based assessments probed depth of understanding rather than rote learning. The teachers appreciated that the students had to use their minds and science skills to solve the problems posed in the assessments. One teacher remarked that what the simulation does, and the paper tests cannot do, is to *animate and show the students the results* of the answer they chose. In addition, teachers saw that the assessments provided information about the processes students were using, not just the answer. Teachers observed that the assessments presented authentic problems that allowed students to see how the science they were studying related to real life and answered the question, "Why do we care?". Several teachers suggested that the simulations would help English language learners and students who didn't do well on traditional tests. All of the teachers said that the most useful information would be score reports on their students' progress and difficulties related to the specific content and inquiry skills. Teachers felt that real-time scoring and immediate results would help them decide what to do next with students. Four of the teachers remarked that they would like to use the simulations for instruction and to administer them during a unit as formative assessments.

7.2 Evidence from student response processes

7.2.1 Feasibility and usability testing

Feasibility and usability testing was conducted through cognitive labs for each of the two force and motion and the two ecosystem assessments with at least five middle school and high school students. Feasibility related to the logistics of the administration and whether the intended constructs were actually elicited by tasks and items. 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. Usability related to the students' ability to navigate through the computer-based assessments and to manipulate the user interfaces. These initial cognitive labs contributed preliminary evidence on the construct validity of the Calipers tasks and items. Only minor revisions were required.

7.2.2 Pilot test responses

Students' responses to the selected response and technology-based items on the pilot test assessments were scored by the computer system, but the constructed responses were scored by human raters. Raters participated in training sessions prior to scoring each item. For each item, raters first discussed the rubric and scored approximately five papers together. Raters then completed approximately six to seven 'calibration' papers. Each calibration paper was scored by raters 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. Inter-rater 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. These inter-rater data supported the technical quality of the constructed response items.

7.2.3 Item level characteristics

Technical quality of individual items was judged using classical test theory and item response theory measures. Across the force and motion and the ecosystem assessments, the p-values averaged 0.65 and ranged from a low of 0.22 to a high of 0.97. Three items in Mountain Rescue were very easy with over 90% of the students giving a correct response and one item had a negative point biserial, which means that it was not distinguishing well between high and low student performances. The distribution of scores for the constructed items that were scored using rubrics generally demonstrated a good spread of responses, but with a slight skew to the upper and lower range of the scoring scale in *Mountain Rescue* assessments, and a skew to the lower range scores in the *FW* assessments. A partial credit IRT model was fitted to the Mountain Rescue data (n = 109) and to the *FW 1* (n = 81) and *FW 2* (n = 83) data. Looking at the weighted mean square fit of the items, all of them fit adequately to the item response model, meaning that they were all contributing to the measurement of the content being tested. In *Mountain Rescue* and *FW 1*, the spread of the difficulty of the items reflected the range of abilities of the student sample population that responded to the items, which means that the items were well-matched to the population. In *FW 2*, however, 12 of the 31 items had item difficulty estimates that were beyond the ability level of the most able student, which means that those items were challenging to all students in the sample. In summary, the technical quality of the individual assessment items that were in the simulation-based assessments were well within acceptable psychometric ranges, although the sample of high school students who worked with FW 2 found the items fairly difficult, which corresponds with the fact that the survey of the opportunity to learn the topics showed that this group of students had less opportunity to learn this content.

7.2.4 Reliability

The reliability of the assessments ranged from .70 to .91 (Cronbach's alpha), which is within accepted usual ranges of reliability for assessments that contain a mix of auto-scored, selected response and human-scored constructed response items.

7.3 Evidence based on the internal structure of the assessments

7.3.1 Evidence-centred design

The Calipers project used evidence-centred assessment design methods to produce reusable 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 and Haertel, 2006). Specifications for the particular 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.

7.3.2 Validity of the content and science inquiry constructs

To examine how well the Calipers items could detect different aspects of the science content and different strands of science inquiry skills, we analysed the data for FW 1 and 2 using a multidimensional partial credit IRT model. This analysis accomplished two things. First, it allowed us to compare the earlier analyses that assumed that content and inquiry skills comprised a single dimension, which can be thought of as 'science knowledge and skills' in the domain. Second, it showed how well the individual content and science inquiry dimensions were being measured in the Calipers items. Results of theses analyses are presented and discussed below.

7.3.3 Content dimensions

The content of items in the ecosystems assessments were coded to particular subdimensions of science content. Through an IRT partial credit analysis that analysed the items as fitting to these sub-dimensions, we were able to show how well items fit to that model. Table 6 presents a comparison of the unidimensional IRT model fit with the multidimensional model fit for both FW 1 and FW 2. The table reports the deviance of each model, which can be interpreted as a measure of how well the IRT model fits the data analysed. The lower the deviance, the better the fit of the model. As can be seen in both cases, the multidimensional model that took account of the fact that items can be coded to different content dimensions fit better for both FW 1 and 2. The differences in the fit were shown to be statistically significant too, using a chi-square test with two degrees of freedom. This means that the assessment items in Calipers can be effectively used to measure different dimensions of science content, and that this approach will yield a more accurate measure of student ability than treating the science content knowledge as unidimensional (i.e., not distinguishing among the component parts of knowledge that make up a student's understanding of the FW ecosystem).

Table 6 Comparison of IRT model fit for unidimensional and multidimensional models in FW on content categories (n = 81)

	Deviance (model fit)		*Difference in model fit*
	Undimensional model (54 parameters estimated)	*Multidimensional model (67 parameters estimated)*	
FW 1	2,892.88	2,342.21	640.67*
FW 2	3,382.85	2,103.75	1,280.10*

Note: *p < 0.05, df2

7.3.4 Science inquiry dimensions

The same method used to judge if the Calipers assessments were suitable for detecting different dimensions of science content was also applied to investigate whether or not the assessments were able to effectively detect various dimensions of students' science inquiry skills. Particular items were coded to address different science inquiry skills. For the analyses, the items were coded and analysed according to the science inquiry skills they were judged to assess. Table 7 shows the results of the analyses. The fit of the multidimensional models that took account of the students' science inquiry skills across sets of items was significantly better than the unidimensional models that did not take account of different dimensions of inquiry. This demonstrates that the Calipers assessments were effective in measuring various inquiry skills as students use the simulation-based assessments.

Table 7 Comparison of IRT model fit for unidimensional and multidimensional models in FW on inquiry categories (n = 81)

	Deviance (model fit)		*Difference in model fit*
	Undimensional model (54 parameters estimated)	*Multidimensional model (67 parameters estimated)*	
FW 1	2,892.88	2,483.09	499.79*
FW 2	3,382.85	2,827.21	556.64*

Note: *p < 0.05, df2

7.4 Evidence from relations to other variables

To provide discriminant validity evidence, we examined how the student performance on the assessments varied in relation to teachers' own assessments of their students' overall ability in science. Teachers were asked to rank their students as low, medium or high achievers in science. All constructed response items were recorded on a scale from 0–1 for the purpose of these analyses. For force and motion, since students had so little opportunity to learn the curved path concepts, items related to curved path key ideas were not included in the analyses of student performance. In addition, for force and motion, findings on the selected response item related to key idea 2 on balanced forces are not reported because of significant missing data for this item.

Figure 4 Graph of mean standardised content score by student achievement ranking of students by teacher

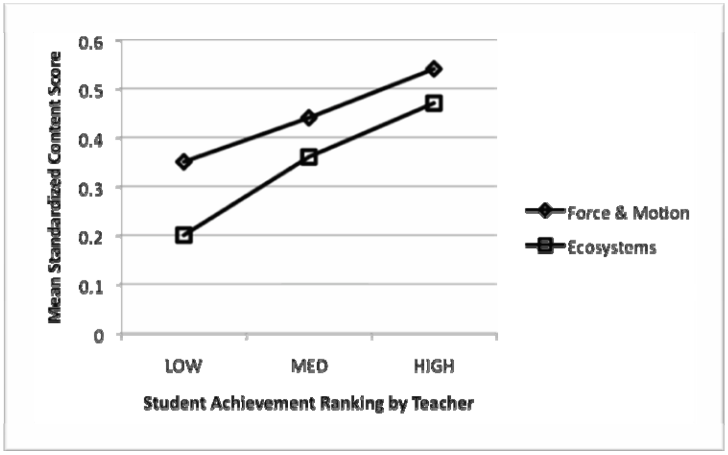
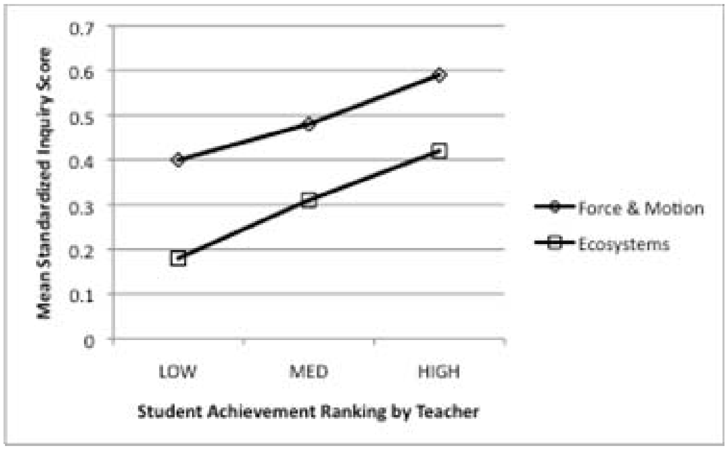


Figure 5 Graph of mean standardised inquiry score by student achievement ranking of students by teacher



Figures 4 and 5 show how the mean scores for science content and inquiry on the simulation-based assessments for force and motion and ecosystems varied based on these rankings. As expected, for both science content and for inquiry in each assessment, the group of students ranked by their teachers as low achievers scored lowest on average, medium achievers scored next highest and the high achievers scored highest. This is evidence that, for measuring science content and inquiry, the assessments do effectively distinguish between levels of performance that are related to overall achievement in science at school.

7.5 Summary of findings on the technical quality, usability and feasibility of the simulation-based assessments

The range of evidence collected indicated that the simulation-based assessments were reliable and valid. The intended alignment with national science standards for the content and inquiry abilities that resulted from the systematic development process was confirmed by independent experts and teacher reviews of the assessments. Pilot testing of the assessments with middle school students showed that the items performed within commonly accepted levels on standard psychometric measures of item difficulty and fit. Multidimensional analysis showed that the assessments did detect science content knowledge and inquiry ability constructs as they were designed to, and the assessments were shown to distinguish clearly between different categories of students based on an external measure of science achievement. The classroom pilot testing provided strong evidence of the assessments' quality and utility as benchmark assessments to test end-of unit achievement of the targeted complex science learning. The pilot test results also provided evidence that the particular exemplars and others that could be modelled after them would be likely to provide credible data for summative accountability purposes.

8 Measuring inquiry skills through student interactions with the simulations

In several parts of the simulation-based assessments, students interacted with the simulation to choose variables, manipulate their values, and run trials using a model. One of the promises of simulation-based assessments is that the actions that students take in their interactions will allow measurement of particular inquiry skills, such as designing and conducting investigations and gathering, analysing and interpreting data. To investigate these possibilities, we analysed data from the student interactions within sections of the FW 2 assessment.

First, we correlated student ability, as measured by performance on all the scored items in the FW 2 assessment, as the dependent variable, with the other non-scored variables of interest in those items where students interacted with the simulation to see what relationships existed between overall performance and those measures. The best significant correlations between non-scored variables in various items and ability were seen on the *number of trials* (significant correlations ranged from .44 to 0.45, $p > .01$); on *correct trials* (significant correlations of .43, $p > .01$); and on *time spent* on the task (significant correlations ranged from .37 to .54, $p > .01$).

Next, we used stepwise regression analyses to explore how much variance in predicted ability was due to non-score variables such as number of trials, number of correct trials and time. An independent variable was added to the regression model as long as its addition contributed a positive increase in the R-square value of the model (the percentage of variance in ability that is predicted by the variables included in the model for that item). Table 8 shows the results of the regression analyses for five items that included non-score measures. The non-score variables measured all contribute significantly to the regression models, in varying amounts. For the simulation competency (SimComp) question, an item that tested students' ability to use the population model, we created a variable that represented the number of correct scores divided by time, and this proved to explain a significant part of the variation in predicted ability. For question 26 in which students were asked to develop a stable ecosystem, the number of trials in which the population died out proved to be significant but as a negative value, meaning that the greater the number of a student's unsuccessful trials, the less likely the student was to be a higher performing student overall. These results indicate that the non-scored variables are worth measuring in addition to the traditionally scored variables. We have yet to fully investigate the most effective ways to combine the two types of scores to maximise the information about the students. This will be part of our ongoing work.

Table 8 The best stepwise models for technology-based items (excluding scores)

Variables	*B*	*SE B*	*β*
SimComp			
SimComp score/time	1.2	0.31	0.53**
Question 17			
Q17 correct inputs	0.97	0.32	0.43**
Q17 correct trials	0.78	0.37	0.29*
Question 19			
Q19 correct trials	1.29	0.48	0.43*
Question 21			
Q21 number of trials	0.2	0.06	0.46**
Question 26			
Q26 number of trials	0.95	0.36	0.36*
Q26 correct input	0.8	0.24	0.47**
Q26 number of tries die out	–1.16	0.38	–.42**

Notes: R^2 = .28 for SimComp; R^2 = .31 for Q17; R^2 = .18 for Q19; R^2 = .22 for Q21; R^2 = .47 for Q26. *p < .05; **p < .01.

9 The promise of simulation-based science assessments

The Calipers demonstration project aimed to provide evidence of the technical quality, feasibility and utility of simulation-based science assessments. The assessments of middle and secondary level science standards for force and motion and ecosystems were designed according to principles of evidence-centred design, developed according to explicit design specifications and subjected to rigorous, iterative review and revision. This systematic approach resulted in assessments that were reliable within accepted standards for assessments that incorporate a mixture of selected response and constructed response items and were shown to be valid for their intended purpose as end-of-unit benchmarks.

The principled design and development processes forged in this project will provide models for development of other simulation-based assessments. The environments modelling scientifically-based principles will allow the use and re-use of the underlying rules 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 were used to design items testing factual content as well as interrelated knowledge of systems. Inquiry tasks asking students to design, conduct, analyse and interpret data and communicate findings were developed. These simulation environments developed for fundamental science systems can be re-used to design assessments of a broader range of science standards for the 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.

The promise of simulation-based science assessments is being further studied in WestEd's SimScientists program. Projects are applying the designs and processes to additional system models in life, physical and earth science. Curriculum-embedded formative assessments are under development that aim to inform and improve student science learning by adding immediate, individualised feedback and customised, graduated coaching. Science simulations are also being developed and evaluated as curriculum supplements. Moreover, the SimScientists program is conducting research on the role simulation-based assessments can play in balanced state science assessment systems. The powerful capabilities of simulations can permit assessment of knowledge and standards not well measured by paper-based formats. The development of systematically designed science simulations promises to revolutionise both instruction and assessment.

References

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

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

Buckley, B.C., Gobert, J.D., Kindfield, A.C.H., Horwitz, P., Tinker, R.F., Gerlits, B., Wilensky, U., Dede, C. and 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*, Vol. 13, pp.23–41.

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

Hickey, D.T., Kindfield, A.C.H., Horwitz, P. and Christie, M.A.T. (2003) 'Integrating curriculum, instruction, assessment and evaluation in a technology-supported genetics learning environment', *American Educational Research Journal*, Vol. 40, pp.495–538.

Horwitz, P., Gobert, J., Buckley, B.C. and Wilensky, U. (2007) *Modeling Across the Curriculum Annual Report to NSF: The Concord Consortium*.

Jackson, S., Stratford, S., Krajcik, J. and 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. and Fishman, B. (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. and Haertel, G. (2006) 'Toward a psychology of educational achievement testing: practical guidelines and principles', *Educational Measurement: Issues and Practice*, Winter, pp.6–20.

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. and Wilson, M. (2003) 'Design patterns for assessing science inquiry', PADI Technical Report 1, SRI International, Center for Technology in Learning, Menlo Park, CA.

Pellegrino, J., Chudowsky, N. and Glaser, R. (2001) *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academy Press, Washington, DC.

Quellmalz, E.S. and Haertel, G. (2004) '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., DeBarger, A.H. and 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, SRI International, Menlo Park, CA.

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

White, B.Y. and Frederiksen, J.R. (1998) 'Inquiry, modeling and metacognition: making science accessible to all students', *Cognition and Instruction*, Vol. 16, pp.3–118.