

*Research Article***Science Assessments for All: Integrating Science Simulations Into
Balanced State Science Assessment Systems**Edys S. Quellmalz,¹ Michael J. Timms,² Matt D. Silbergliitt,³ and Barbara C. Buckley¹¹*WestEd, 400 Seaport Court, Redwood City, California 94063*²*Australian Council for Educational Research, Camberwell, Victoria, Australia*³*WestEd, 300 Lakeside Drive, Oakland, California 94612**Received 16 June 2011; Accepted 17 December 2011*

Abstract: This article reports on the collaboration of six states to study how simulation-based science assessments can become transformative components of multi-level, balanced state science assessment systems. The project studied the psychometric quality, feasibility, and utility of simulation-based science assessments designed to serve formative purposes during a unit and to provide summative evidence of end-of-unit proficiencies. The frameworks of evidence-centered assessment design and model-based learning shaped the specifications for the assessments. The simulations provided the three most common forms of accommodations in state testing programs: audio recording of text, screen magnification, and support for extended time. The SimScientists program at WestEd developed simulation-based, curriculum-embedded, and unit benchmark assessments for two middle school topics, Ecosystems and Force & Motion. These were field-tested in three states. Data included student characteristics, responses to the assessments, cognitive labs, classroom observations, and teacher surveys and interviews. UCLA CRESST conducted an evaluation of the implementation. Feasibility and utility were examined in classroom observations, teacher surveys and interviews, and by the six-state Design Panel. Technical quality data included AAAS reviews of the items' alignment with standards and quality of the science, cognitive labs, and assessment data. Student data were analyzed using multidimensional Item Response Theory (IRT) methods. IRT analyses demonstrated the high psychometric quality (reliability and validity) of the assessments and their discrimination between content knowledge and inquiry practices. Students performed better on the interactive, simulation-based assessments than on the static, conventional items in the posttest. Importantly, gaps between performance of the general population and English language learners and students with disabilities were considerably smaller on the simulation-based assessments than on the posttests. The Design Panel participated in development of two models for integrating science simulations into a balanced state science assessment system. © 2012 Wiley Periodicals, Inc. *J Res Sci Teach* 49: 363–393, 2012

Keywords: assessment; simulations; model-based learning; balanced assessment systems; evidence-centered design

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: US Department of Education; Contract grant number: 09-2713-126; Contract grant sponsor: National Science Foundation; Contract grant number: 0733345.

Correspondence to: E.S. Quellmalz; E-mail: equellm@wested.org

DOI 10.1002/tea.21005

Published online 23 January 2012 in Wiley Online Library (wileyonlinelibrary.com).

State science assessments are engulfed in a sea change. Numerous standards-setting groups recommend that K-12 science education shift its focus to fewer, more integrated core ideas, deeper understanding of dynamic science systems, and the use of science inquiry practices. For example, the recent College Board *Standards for Science Success* and the National Research Council *Framework for Science Education* recommend deeper learning of the fundamental nature and behavior of science systems, along with the inquiry practices scientists use to study system dynamics (College Board, 2009; National Research Council [NRC], 2011). Many states recognize that traditional assessment formats cannot adequately assess these aspects of science. To date, computer technologies have been used mainly to address the logistics of administration and scoring of assessments, but now technologies are beginning to show promise for developing and delivering measures of complex learning useful for instruction and policy (Quellmalz & Pellegrino, 2009).

Technologies can present students with rich task environments that model key features of science systems in action in the natural world. Science simulations can present authentic environments structured according to principles in the domain. Spatial, temporal, and causal phenomena can be represented that may be otherwise unobservable and not directly manipulable because they are too large (hurricanes), or too small (chemical reactions), too fast (earthquakes), or too slow (plant growth). Because simulations are interactive, students can demonstrate their abilities to apply the active inquiry practices of science by designing investigations, conducting iterative trials, predicting, observing, and explaining findings, and critiquing the investigations of others. Simulation-based software can automate and individualize feedback to students, provide immediate, customized coaching, and generate progress and proficiency reports to teachers and students, for individuals and groups.

Simulations are becoming a component of large-scale science assessments. The Programme of International Student Assessment (PISA) has pilot-tested science simulations such as the functions of a nuclear reactor and the exploration of the genetic breeding of plants. More simulation-based science tasks are planned for 2015. The 2009 National Assessment Educational Progress (NAEP) administered Interactive Computer Tasks (ICT) and will field more in the next science assessment. The 2014 NAEP for Technology and Engineering Literacy will also administer interactive scenario-based tasks. At the state level, Minnesota has an online science test with simulated laboratory experiments or investigations of phenomena such as weather or the solar system (Minnesota Department of Education, 2010). Utah is piloting science simulations (King, 2011). The state testing consortia are designing technology-enhanced items to test English Language Arts and Math common core standards. Tests of the forthcoming *Next Generation Science Standards* are likely to include simulations.

The next generation of state assessment systems being developed by state collaboratives aims to achieve balanced, multilevel assessment systems that provide mutually reinforcing information about student achievement gathered from curriculum-embedded, benchmark, and summative assessments that dovetail across classroom, district, and state levels (Darling-Hammond & Pecheone, 2010). At the same time, states must ensure that these technology-based assessments are accessible to the special populations of students required for inclusion in state testing. Technology can provide new supports for including English language learners (ELL) and students with disabilities (SWD) in state assessment systems. A new generation of simulation-based science assessments is showing the potential to transform what, how, when, where, and why assessment occurs and how it can support teaching and learning for all students.

This article presents research findings on the technical credibility and practical suitability of simulation-based science assessments for inclusion as components of a state science

assessment system. The study, “Integrating Simulation-Based Science Assessments into Balanced State Science Assessment Systems,” was funded by the U.S. Department of Education, Office of Elementary and Secondary Education as an Enhanced Assessment Grant (EAG). The study was a collaboration, led by Nevada, involving Connecticut, Massachusetts, North Carolina, Utah, and Vermont; WestEd’s SimScientists program, and the National Center for Research on Evaluation, Standards, & Student Testing (CRESST) at University of California Los Angeles. The study addressed two main questions: (1) Could simulation-based science assessments be developed that were of sufficient psychometric quality, feasibility, and utility to warrant inclusion as components of a state science assessment system? and (2) What models might guide the integration of simulation-based assessments into the state science system?

The SimScientists program at WestEd had developed two suites of simulation-based assessments (Ecosystems and Force & Motion) for use in middle school classrooms as part of the Calipers II project funded by the National Science Foundation. For each topic, simulation-based, curriculum-embedded assessments provided opportunities for classroom-level formative assessment, off-line reflection activities that reinforced and extended the targeted concepts and inquiry skills, and simulation-based unit benchmark assessments that provided summative proficiency data. To increase accessibility for students who needed accommodations, the EAG project added audio and screen magnification accommodations along with support for completing the assessment over multiple class periods. Science leaders from the six states formed a Design Panel to monitor the implementation of the assessments during the field test, review assessment and evaluation findings, and summarize implications for their state science assessment systems. Three of states (Nevada, North Carolina, and Utah) volunteered to field-test the SimScientists assessments in their middle school science classrooms.

Theoretical Frameworks

The study integrated frameworks for assessment and science learning. The assessment framework set forth purposes, roles, and designs of the SimScientists assessments within the context of state assessment systems and utilized contemporary measurement methods for systematically designing, developing, and validating the assessments. The science-learning framework, used within the assessment framework, specified science knowledge and inquiry practices critical to the domain of science. These frameworks were aligned with national and state science standards and informed by learning research.

SimScientists Assessment Framework

Balanced Assessment Systems. The NRC report, *Knowing What Students Know*, advocates a balanced assessment system that relies on a nested system of assessments that exhibits features of comprehensiveness, coherence, and continuity (Pellegrino, Chudowsky, & Glaser, 2001). *Comprehensiveness* is achieved by multiple measures of the full range of standards. *Coherence* involves a *horizontal* alignment of standards, goals, assessments, curriculum, and instruction, as well as *vertical* alignment among assessments at different levels of the assessment system. *Continuity* is achieved by going beyond annual, high-stakes tests to multiple assessments over time and in time for teachers to tailor instruction.

Multiple measures of state standards are particularly important because annual, large-scale tests are alleged to tap only subsets of goals, narrowing and distorting curricula in unintended ways (Darling-Hammond, 2010; Darling-Hammond & Pecheone, 2010). Nor do most large-scale assessments include item types that can measure extended thinking,

reasoning, problem-solving, or inquiry (Quellmalz, DeBarger, Haertel, & Kreikemeier, 2005). Researchers and policy makers are pursuing models for building assessment systems that vertically articulate evidence gathered at classroom, district, and state levels (Quellmalz & Moody, 2004; NRC, 2006; Stiggins, 2006; Perie, Marion, & Gong, 2009). Multilevel assessment systems allow for the use of curriculum-embedded assessments that provide immediate feedback for learning, the incorporation of complex tasks more aligned to the skills involved in 21st century learning, and the generation of student assessment information at all levels of the system, resulting in a rich profile of what students know and can do (Darling-Hammond, 2010; Darling-Hammond & Pecheone, 2010).

The SimScientists assessments for Ecosystems and Force & Motion were designed to supplement state science test evidence by providing science assessments to be (1) embedded within curriculum units that could serve formative assessment purposes by providing immediate feedback, monitoring progress, and informing needed adjustments to instruction, and (2) administered at the end of a unit as summative measures of proficiency on the targeted science content and inquiry practices.

Evidence-Centered Assessment Design. Evidence-centered assessment design (ECD) facilitates assessment coherence by linking the targets to be assessed with evidence of proficiency on them, and with tasks and items eliciting that evidence (Messick, 1994; Mislevy & Haertel, 2007). The process begins by specifying a student model of the knowledge and skills to be assessed. The ECD design process aligns the student model with an evidence model that specifies which student responses are evidence of targeted knowledge and skills, how student performances are to be analyzed, and how they will be reported. The student and evidence model are then aligned with a task model that specifies features of the tasks and questions intended to elicit student performances that provide evidence of the targeted knowledge and skills.

The SimScientists assessments used the evidence-centered design method to align the science content and inquiry to be assessed, to scoring and reporting methods, and to the specification of the assessment tasks and items. Sources shaping the designs of the SimScientists assessments' student, task and evidence models included learning research on the development of expertise, research on science learning, research on the affordances of simulations and technology for learning and assessment, and the consensus of science professional organizations on standards for K-12 science education.

Framework of Science Learning

Model-Based Learning. Across academic and practical domains, research on the development of expertise indicates that experts have acquired large, organized, interconnected knowledge structures, called schema, and well-honed, domain-specific problem-solving strategies (Bransford, Brown, & Cocking, 2000). Jacobsen characterized the schema of experts as "complex systems" mental models in contrast to the deterministic "clock-work" mental models of novices (Jacobson, 2001). A growing body of research shows that scientists use schema composed of physical, mathematical, and conceptual models as tools for generating and testing hypotheses and to communicate about natural and designed systems. Model-based reasoning is a signature practice of the sciences that supports how scientists create insights and extend understandings of nature (Clement, 1989; Nersessian, 2008).

Research on model-based learning suggests that effective science learners also form, use, evaluate, and revise their mental models of phenomena in a recursive process that results in more complete, accurate, and useful mental models of a science system (Buckley, 2000;

Gobert, 2000; Gobert & Buckley, 2000; Gobert & Clement, 1999). For example, students who participate in cycles of model-based reasoning build deeper conceptual understandings of core scientific principles and systems, interpret patterns in data, and formulate general models to explain phenomena (Stewart, Carter, & Passmore, 2005; Lehrer, Schauble, Strom, & Pligge, 2001). Further, cognitive research shows that learners who internalize schema of complex system organization—structure, functions, and emergent behaviors—can transfer this heuristic understanding across systems (e.g., Goldstone, 2006; Goldstone & Wilensky, 2008). These studies informed the SimScientists assessments' focus on measuring student understanding of science systems in terms of their components, interactions, and system behavior, and the science practices used to study them.

Simulations in Science Learning. Using simulations to provide dynamic representations of spatial, temporal, and causal phenomena, scientists can represent their understanding of these phenomena (Hestenes, Wells, & Swackhamer, 1992; Stewart & Golubitsky, 1992) and support schema formation and mental model construction (Norman, 1993). Learners, too, can use simulations to investigate problems that involve phenomena that are too large, too small, too fast, too slow or too dangerous to study in classrooms (Buckley et al., 2004; Vattam et al., 2011). Moreover, simulations can generate and superimpose multiple physical and symbolic representations to reduce potentially confounding language demands (Kopriva, 2008). The SimScientists assessments designed science simulations both to represent models of dynamic science phenomena and to engage students in investigations that would elicit evidence of the system knowledge and inquiry practices.

Learning Research. A series of reports from the National Research Council synthesizes decades of research on human learning and strategies for promoting its development. They include *How People Learn: Brain, Mind, Experience and School* (Bransford et al., 2000), *Knowing What Students Know: The Science and Design of Educational Assessment* (Pellegrino et al., 2001), *How Students Learn History, Mathematics, and Science in the Classroom* (Donovan & Bransford, 2005), and *Taking Science to School* (Duschl, Schweingruber, & Shouse, 2007). The SimScientists projects drew from these reports key principles for designing the assessment tasks.

MEANINGFUL LEARNING. Learning theory holds that the learning environments in which students acquire and demonstrate knowledge should represent contexts of use (Simon, 1980; Collins, Brown, & Newman, 1989). The *Taking Science to School* report recommends that rather than teaching and testing individual skills separately, skills be taught and tested in the context of a larger investigation linked to a driving question.

ACTIVE INVESTIGATIONS. Researchers have found that K-8 students, with appropriate scaffolding, can engage in investigations, make hypotheses, gather evidence, design investigations, evaluate hypotheses in light of evidence, and build their conceptual understanding (Geier et al., 2008; Lehrer & Schauble, 2002; Metz, 2004). Research finds that all students, particularly English language learners, can benefit greatly from inquiry-based science instruction that depends less on mastery of English than do de-contextualized textbook knowledge or direct instruction by the teacher (Lee, 2002). Incorporating scientific argument throughout investigations adds the element of convincing peers of the explanation, responding to critiques, and reaching consensus (Bell & Linn, 2000; Duschl & Osborne, 2002). Often projects include a culminating activity in which students make a presentation or develop a poster to communicate their findings (Edelson & Reiser, 2006; Krajcik, Blumenfeld, Marx, Bass, & Fredricks, 1998; Reiser et al., 2001).

A number of studies provided evidence that these project-based experiences helped students learn science inquiry practices. Kolodner et al. (2003) found that middle school students who practiced inquiry in several project-based science units performed better on the inquiry tasks of scientific practice (as measured by performance assessments, Quellmalz, Schank, Hinojosa, & Padilla, 1999) than students from traditional classrooms. Assessments like the NAEP, state tests, and items from the Trends in International Mathematics and Science Study (TIMSS) administer scenario-based sets of inquiry tasks (Marx et al., 2004; Rivet & Krajcik, 2004). Simulations have been recommended for inclusion in state science tests (National Research Council, 2006; Quellmalz & Haertel, 2004). The SimScientists curriculum-embedded formative assessments and the summative benchmark assessments integrated measurement of conceptual understanding as students engaged in science inquiry practices.

SCAFFOLDING. Students benefit from “scaffolds” that embed instructional guidance in ongoing investigations (Linn, Bell, & Davis, 2004; Quintana et al., 2004; Reiser, 2004). Combinations of graphics and verbal descriptions can highlight key concepts and procedures (Pashler et al., 2007). Process scaffolds can cue important components of inquiry or science arguments (Duschl et al., 2007). The SimScientists curriculum-embedded formative assessments employ such scaffolding techniques.

FORMATIVE ASSESSMENT. Formative assessments play a critical role in balanced assessment systems. Formative assessments combine gathering evidence of learning progress with scaffolding that functions as additional differentiated, individualized instruction. Effective formative assessment provides “short term feedback so that obstacles can be identified and tackled” (Black, 1998, p. 25) and is an important strategy for improving student learning, particularly for low-ability students. The effectiveness of formative assessment depends on several factors including the quality of feedback provided to students, involvement of students in self-reflection and improvement, and whether adjustments are actually made *during* instruction based on the assessments (Black & Wiliam, 1998). Contingent feedback and follow-up instruction that include explanations and worked examples have been shown to promote student achievement (Bangert-Downs, Kulik, Kulik, & Morgan, 1991; Dassa, Vazquez-Abad, & Ajar, 1993; Pashler et al., 2007). Effective feedback includes strategies such as eliciting multiple responses to the same question, asking for evidence to support predictions and explanations, asking for comparisons of ideas and predictions with those of other students, providing evidence of a principle or concept previously discussed or presented, and making connections to other ideas and concepts from prior investigations (Herman, Osmundson, Ayala, Schneider, & Timms, 2005). A study of three teachers who used WISE on-line activities improved student performance over the course of 3 years using evidence provided by the system to change the formative questions teachers asked, add hands-on activities, and modify their teaching strategies (Gerard, Spitulnik, & Linn, 2010). “On-the-fly” assessment by the teacher, assessment conversations, and curriculum-embedded assessments are all acknowledged as effective, research-based strategies for guiding science instruction (Duschl et al., 2007). The SimScientists curriculum-embedded assessments were designed to provide these features of effective formative assessment—ongoing collection of evidence of learning progress, immediate feedback, and customized scaffolding/coaching.

ARTICULATION AND REFLECTION. Cycles of feedback, revision, and reflection are aspects of metacognition that are critical for students to regulate their own learning (Pashler et al., 2007; White and Frederiksen, 1998). Research suggests that small group discussions and debate can

enhance higher-order thinking and development of scientific argumentation skills (Blumenfeld, Marx, Krajcik, & Soloway, 1996; Vye et al., 1998). The SimScientists curriculum-embedded simulation-based assessments provide immediate feedback, opportunities for revising responses, and tasks engaging students in self-assessment of their predictions, conclusions, and explanations. In addition, each embedded simulation-based assessment is followed by an off-line reflection activity that provides opportunities for scientific discourse, argumentation, and presentations.

Technology Affordances. Simulations allow students to see dynamic science systems “in action” and to actively manipulate the system interactions and behaviors. Simulations can present simple, grade-appropriate models of the science system components, interactions, and emergent system behaviors. Multiple representational formats can juxtapose and overlay concrete and symbolic representations. For example, animations of predator/prey interactions can appear simultaneously alongside population graphs and tables.

Multimedia researchers have examined the effects of pictorial and verbal stimuli in static, animated, and dynamic modalities, as well as the effects of active versus passive learning enabled by degrees of learner control (Mayer, 2005). A great deal of research has been conducted on external forms of *stimulus representations*. Research on the perceptual correspondence of models to the natural systems they represent (e.g., solar system, cells, circuits, ecosystems) suggests features to consider in the design of science assessment tasks. Research on models’ physical similarity to a natural system and the ways in which system interrelationships are depicted through conventional physical and symbolic forms and signaled or highlighted, can inform the design of science assessment tasks. In a review of animation and interactivity principles in multimedia learning, Betrancourt (2005) noted that multimedia representations have evolved from sequential static text and picture frames to increasingly sophisticated visualizations. Animations are considered particularly useful for providing visualizations of dynamic phenomena that are not easily observable in real space and time scales (e.g., plate tectonics, circulatory system).

When degrees of *learner control and interactivity* are introduced as variables, other research suggests that spatial representations enable effective mental simulations and visualizations (Schwartz & Heiser, 2006). In the *active modality*, learners can go beyond passive views of dynamic stimuli by controlling the pacing and direction of an animation. Rebetez, Sangin, Betrancourt, and Dillenbourg (2004) found that the active form of learner-control of continuous animation led to better comprehension than a succession of static snapshots. Animations become *interactive* simulations if learners can manipulate parameters as they generate hypotheses, test them, and see the outcomes, therefore taking advantage of technological capabilities well suited to conducting scientific inquiry. Rieber, Tzeng, and Tribble (2004) found that students given graphical feedback during a simulation on laws of motion with short explanations far outperformed those given only textual information. An important finding is that individuals’ initial differences in spatial reasoning ability tend not to make a difference when well-structured spatial representations are presented (Heiser, 2004).

A review of research studies focusing on design features that affect student learning with science simulations identified sets of features that clustered around effective interfaces, powerful visualizations, and illuminating inquiry (Scalise et al., 2011). Visual focal points for substantive content, sensitivity to cognitive load, using dynamic representations and abstractions mindfully, and recommendations for designing inquiry with simulations are particularly salient to the design of simulation-based assessments. The SimScientists assessments

referenced this and other multimedia design research in the design of the simulation environments, the tasks and questions, and the types of student responses.

Accessibility for English Language Learners and Students With Disabilities. Building on work by Rose and Meyer (2000), CAST (2008) developed a framework for Universal Design for Learning (UDL) recommending three kinds of flexibility: (1) representing information in multiple formats and media, (2) providing multiple pathways for students' action and expression, and (3) providing multiple ways to engage students' interest and motivation. Assessments, too, should present information in more than one modality (e.g., auditory and visual), allow simultaneous presentation of multiple representations (e.g., scenes and graphs), and vary simple and complex versions of phenomena and models. Multiple pathways for expression may include interactivity, hints and worked examples, and multiple response formats (drawing, writing, dragging, and dropping).

Universal design for computer-based testing (UD-CBT) further specifies how digital technologies can create tests that more accurately assess students with a diverse range of physical, sensory, and cognitive abilities and challenges through the use of accommodations (Burling et al., 2006; Harns, Burling, Hanna, & Dolan, 2006). Accommodations are defined as changes in format, response, setting, timing, or scheduling that do not alter in any significant way what the test measures or the comparability of scores (Phillips, 1993). UD-CBT has been found to level the playing field for English language learners (ELL) and students with disabilities (Case, Brooks, Wang, & Young, 2005; Wang, 2005). Tools already built into students' computers can allow multiple representations (text, video, audio), multiple media, highlighters, and screen magnification (Case, 2008; Twing & Dolan, 2008). The SimScientists assessments designs drew on these studies to design accommodations commonly allowed in state testing programs— audio recordings of text, screen magnification, and segmentation to support re-entry at the beginning of a task if extended time is needed.

SimScientists Assessment Design Principles

The design principles that shaped the SimScientists assessments emerged from integrating the assessment and science frameworks. The assessment framework first laid out the roles for the simulation-based assessments in a state science assessment system. Formative purposes were identified for the curriculum-embedded assessments. Unit benchmark assessments were specified to serve interim summative purposes.

Evidence-centered assessment design was employed to closely link the student, task, and evidence models of the simulation-based formative and summative assessments. The *student model*, specifying the content and inquiry to be tested, was shaped by the system model of components, interactions, and emergent behaviors, and associated inquiry practices, as well as by national and state science frameworks. The *task model* specified the nature of the science system representations and the response demands of the assessment tasks and items that would elicit evidence of the targeted science content and inquiry. These assessment tasks presented authentic, problem-based inquiry. For the curriculum-embedded formative assessments, the system provided scaffolds for conceptual understanding and inquiry processes in the form of principled formative feedback and coaching cycles. The reflection activities provided opportunities for sense-making and scientific discourse as well as reflection and self-assessment that support metacognitive self-regulation. The summative unit benchmark assessments did not provide feedback and coaching. The *evidence model* specified the student responses that would constitute evidence of proficiency on the targets, the ways in which the

responses would be evaluated and scored, and how the scores would be summarized to characterize proficiency levels.

Description of SimScientists Assessments

The SimScientists assessments represent a shift from testing discrete factual content to a focus on connected knowledge structures that organize concepts and principles into crosscutting features of all systems—components, interactions, and emergent behaviors—and the inquiry practices used to investigate them. SimScientists assessment suites are composed of two or three embedded formative assessments that the teacher inserts into a unit at key points and a summative benchmark assessment at the end of the unit. The SimScientists Learning Management System (LMS) delivers the assessments and collects data over the Internet. The LMS enables teachers to assign assessments, view progress reports, assign differentiated follow-up classroom reflection activities, score constructed responses from the benchmark assessment, and view summative proficiency reports.

The tasks and items designed for the simulation-based assessments made use of the flexibility provided by digital technologies to improve access for ELL and SWD populations. SimScientists assessments include on-screen analogs of the three most commonly requested accommodations: (1) the assessments are programmed so that students can stop and restart without losing their data or their place, permitting an extended time accommodation, (2) the assessments can be used with a screen magnification feature that provides an analog of large print, and (3) the assessments can be used with audio files that provide a read-aloud accommodation. All have been shown to benefit ELL and SWD populations (Chiu & Pearson, 1999; Bolt & Thurlow, 2004).

Multiple modes of representation are helpful for ELL and SWD populations. Simultaneous, linked representations also support development of multilevel mental models (Buckley, Gobert, Horwitz, & O'Dwyer, 2010; Horwitz, Gobert, Buckley, & O'Dwyer, 2010; Mayer & Anderson, 1992). Figure 1 provides an example from the population model simulation for ecosystems. In the task shown, students choose the starting values of one or more organisms, and observe a set of icons that represent the population as it grows, declines, or reaches equilibrium. Students also observe a population curve as it is generated, and use a tool called

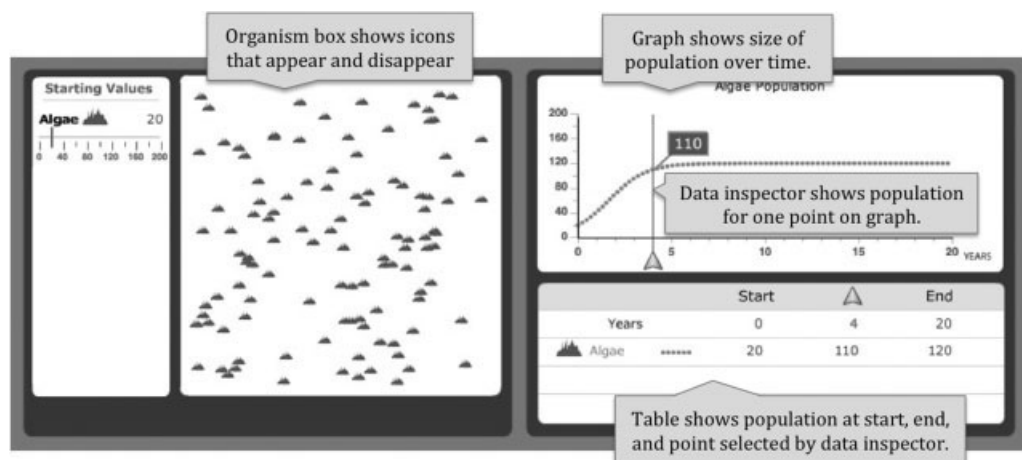


Figure 1. Task model example—multiple modes of representation.

the “data inspector” to find the value of the population at specific points on the curve. These synchronized representations link the structure, functions, and emergent behaviors of the systems being simulated. The center box represents the interactions of component organisms that produce the population changes over time, which are displayed simultaneously in the graph and data table on the right. The interactivity of this simulation gives students control over the representations and enables students to demonstrate inquiry practices as they make predictions, design experiments to test their predictions, interpret data, draw conclusions, and evaluate their predictions.

Each of the assessment suites contained *embedded (formative) assessments* (two in Ecosystems and three in Force & Motion) that were inserted into instruction when the teacher deemed the prerequisites complete. During the embedded assessments, students completed tasks such as making observations, running trials in an experiment, interpreting data, making predictions, and explaining results. They used various methods such as selecting from a choice of responses, changing the values of variables in the simulation, drawing arrows to represent interactions in a system, and typing explanations to complete these tasks. For all but the typed responses, the assessments gave students feedback and graduated levels of coaching so that students had multiple opportunities to correct their errors and confront their misconceptions, with increasing scaffolding based on the amount of help needed. For typed responses, students were given opportunities first revise their response based on criteria (a student-friendly version of a rubric) and then self-assess their revised response by comparing it to a sample answer. Figure 2 presents screenshots of two SimScientists embedded assessments that provided immediate feedback and coaching as students interacted with the simulations.

In the left screenshot, students are asked to draw a food web showing the transfer of matter and energy between organisms based on prior observations made of feeding behaviors in the novel ecosystem. When a student draws an incorrect arrow, a feedback box coaches students to observe again by reviewing the animation and to draw the arrow from the food source to the consumer. Feedback also addresses common misconceptions. Because the assessments capture the values and variables students select during investigations, SimScientists assessments are able to provide coaching for inquiry practices, too. The right screen shot shows feedback and coaching for an investigation of population changes.

Additional feedback was provided in the form of reports to students and teachers at the end of each assessment. Because reports in the form of grades are not as productive and can undermine learning and student motivation, these progress reports provided the kinds of descriptive feedback that helps students connect their success in the assessment to their effort (Covington, 1999; Maehr & Midgley, 1996). Based upon the amount of coaching students

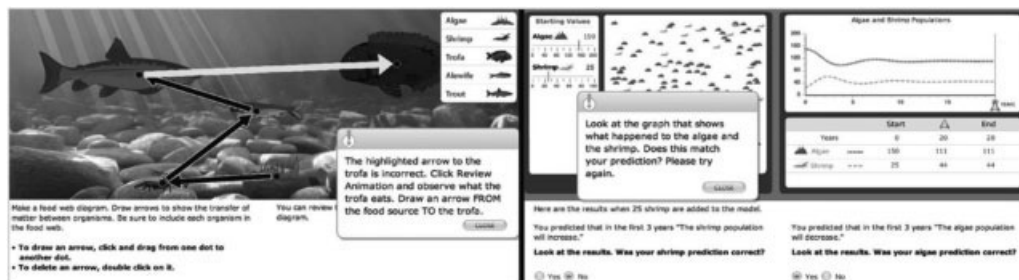


Figure 2. SimScientists embedded assessments provide feedback and coaching.

needed to complete the assessment, the LMS generated a progress report that indicated whether a student is “On Track, Making Progress, or Needs Help” for each content and inquiry target. The progress reports signal the teacher to adjust instruction during subsequent reflection activities. Examples of the progress reports are available in the Supporting Information.

Opportunities for reflection and improvement were provided in classroom reflection activities that followed each of the embedded formative assessments. For example, in the Ecosystem suites, reflection activities stress the big idea that all ecosystems share the same organizational structure and that similar behaviors (e.g., population changes) emerge from this structure. Groups engaged in scientific discourse in order to transfer their science content knowledge and inquiry skills to three new ecosystems (Savanna, Galapagos, Tundra) and prepare presentations that were evaluated by both students and teachers. Detailed reports from the assessment, coupled with the reflection activities, provided teachers with the tools they needed to adjust instruction based on the results of the assessments.

At the end of the curriculum unit, students completed the *benchmark assessment*, which consisted of tasks and items parallel to those in the embedded assessments, but transferred into a new context. For example, the embedded assessments for the ecosystems suite were set in a lake ecosystem (see Figure 1); the benchmark assessment used the same activities, but the setting was a grasslands ecosystem with different organisms and different, although parallel, interactions. For Force & Motion, the embedded assessments presented tasks for a fire truck, while the benchmark assessment represented forces on a train. In this way, students could not simply memorize the material from the embedded assessments, and had to show that they could transfer their knowledge and inquiry practices. No coaching was provided in the summative benchmark. Upon completion of the benchmark assessment, the teacher used the LMS to score students’ written responses using a rubric specified by the assessment designers. These scores, along with the scores from machine-scored tasks, were evaluated by the LMS using a Bayes Net to produce summative proficiency reports to both students and the teacher on the relevant state science standards and specific content and inquiry targets addressed. The benchmark assessment report classifies an individual’s proficiency level (Below Basic [BB], Basic [B], Proficient [P], Advanced [A]) for the content categories (roles, interactions, populations) and on the inquiry targets, (e.g., design, conduct, evaluate). The report also provides a class-level report on the content and inquiry proficiencies. The generation of this report is described in the methods section.

Design of SimScientists Assessments

As outlined in the AERA, APA, and NCME (1999) testing standards, our methodology for test construction and revision followed a process of alignment, quality review, cognitive labs, feasibility testing, pilot testing, and validity studies. Within this overarching framework, evidence-centered assessment design and model-based learning guided the specifications of the student, task, and evidence models.

Alignment

Alignment is a recursive process involving (1) domain analysis to specify the complex science systems and their levels, (2) analyses of national standards appropriate for middle school (National Assessment Governing Board [NAGB], 2008; NRC, 1996; AAAS, 1993), (3) analysis of existing curriculum materials to determine the concepts, representations, vocabulary, and tasks commonly used in classrooms, and (4) a review of the literature to identify common misconceptions.


Model Level	Model Level Descriptions	Content Targets by Model Level	Science Practices by Model Level
 <p>Component</p>	What are the components of the system and their rules of behavior?	Every ecosystem has a similar pattern of organization with respect to the roles (producers, consumers, and decomposers) that organisms play in the movement of energy and matter through the system.	Identify and use scientific principles to distinguish among components
 <p>Interaction</p>	How do the the individual components interact?	Matter and energy flow through the ecosystem as individual organisms participate in feeding relationships within an ecosystem.	Predict, observe, and describe interactions among components.
 <p>Emergent</p>	What is the overall behavior or property of the system that results from many interactions following specific rules?	Interactions among organisms and among organisms and the ecosystem's nonliving features cause the populations of the different organisms to change over time.	Predict, observe, and investigate changes to a system. Explain changes to a system using knowledge about the interactions among its components.

Figure 3. Student model for ecosystems, including model levels, content targets, and inquiry practices.

Student Model

Based on these analyses, the design team drafted student models identifying the components of the systems (including structure and behavior of each component), interactions among components, and behavior and properties that emerge from those interactions. The system model for Ecosystems (shown in Figure 3) presents these three levels of a system, applied to content standards for middle school ecosystems and associated inquiry practices.

The first two columns describe the generic system model levels—components, interactions, and emergent behavior. The third column describes the model levels and more specific content targets for ecosystems. The last column includes inquiry targets for each level. Expert reviews by AAAS and the advisory panel ensured scientific accuracy and appropriateness of the student model for middle school.

Once student models were identified, the team began a recursive process of articulating the context in which to embed the target models, appropriate problem types, the simulations and other representations that students would manipulate and interpret, and the overall sequence of tasks. The designs specified a driving question, a scenario in which the question is to be investigated, and a sequence of tasks that involve increasingly complex models of the phenomena and reasoning during inquiry.

Task Model

Task design focused on inquiry practices identified in NAEP 2009, such as making observations to identify components and interactions, making predictions, designing experiments to test those predictions, interpreting data, evaluating predictions, and explaining results. Task designs also specified what data were to be collected by the system. This included not only students' answers, but also elements of their interactions with the simulation such as variables manipulated in the simulation (e.g., initial population levels), values assigned to variables in the simulation (e.g., number of shrimp), and the number of trials run in an experiment. These types of data cannot be collected in conventional paper and pencil tests. For each assessment the design team developed algorithms for classifying student responses and actions to identify types of errors or common misconceptions. For the embedded formative assessments, the design team also scripted coaching tailored to the types of errors or misconceptions demonstrated.

Evidence Model

The design team specified the data to be collected during student interactions with the simulation-based tasks and coded that data to specific content and inquiry targets. The data included observable events such as answers to questions, inputs to simulations, the full text of constructed responses, and arrows drawn. The design team also specified the algorithms for classifying these data into error classes and for triggering principled levels of coaching (embedded only). The design team further specified the scoring algorithms for the responses that could be automatically scored and provided rubrics and examples for teacher scoring of the constructed responses (benchmark only).

Based on these algorithms, progress reports generated by the embedded assessments classified student performance as “Needs Help, Making Progress, or On Track” for each content and inquiry target. Data collected during the benchmark assessment at the end of the instructional unit were analyzed using a Bayes Net scoring system (see Martin & VanLehn, 1995 and Mislevy, 1994 for examples of the use of Bayes Nets in assessment). The Bayes Net provided estimates of student proficiency on the content and inquiry targets in each model level. The LMS then generated reports for each student and the class.

Storyboard Design

All of these assessment design models were brought together to create the benchmark storyboard that is the most detailed specification for programmers of the student, task, and evidence models of evidence-centered design. It specifies screen by screen what students will see, how they will be able to interact with it, algorithms for analyzing student responses (coded by content and inquiry target), how the system will respond to student actions and answers, what data are sent to the database for later analysis, and what will go into the reports. Storyboards for the embedded assessments also specify the feedback and coaching rules and scripts. Universal Design for Learning (UDL) guidelines dictated such features as the colors and patterns used in graphs or legends used in animations of feeding behaviors. Research in learning with multimedia described earlier was used to consider such factors as cognitive load and visual focal points.

Quality Review

The Design Panel and content experts at AAAS reviewed the storyboards and alignment documents, rating the alignment of the questions and tasks posed with the concepts and inquiry practices of the 2009 NAEP and the AAAS Benchmarks and key ideas. Reviewers provided ratings and recommendations based on the following features:

- Appropriateness of the science system models and content and inquiry targets
- Alignment of assessment targets, tasks and items with the *NAEP 2009*, *AAAS Benchmarks*, and *NSES* science standards.
- Coherence of the assessments: curriculum-embedded, unit benchmark
- Grade-level appropriateness of assessment tasks and items
- Scientific and item quality of the assessments

Usability Testing

The simulation-based assessments were programmed and quality-assurance tested to determine if they performed as designed. Any problems were corrected before we conducted

usability testing. Following protocols employed in the SimScientists projects and other research (Forsyth & Lessler, 2004; Nolin & Chandler, 1996; Zucker, Sassman & Case, 2004), 10 middle school students and 4 teachers participated in cognitive laboratory sessions to provide preliminary evidence of usability and construct validity. In the student cognitive laboratory sessions, each individual student completed one module of the SimScientists assessments. The session was conducted by a trained WestEd researcher who followed a protocol in which the student first practiced verbally “thinking aloud” before completing the SimScientists module. As they did so, screen capture software simultaneously recorded the student’s activities in the assessment module and the audio capture of their spoken thoughts as they completed the tasks. The researcher also used a protocol in which a written record was kept of particular actions that were coded to the content and inquiry targets, for each screen of the assessment. From the written records, counts of the times that students accessed relevant science knowledge and inquiry practices were obtained and problematic tasks, screens, or wording were flagged for potential revision. In the teacher cognitive laboratory sessions, a similar protocol was followed, but teachers were asked to comment on how they thought their students would fare in responding to the assessments. These early trials were used to identify problems with clarity, logistics, and usability, and as early indicators of construct validity, that is, that the tasks are eliciting the intended knowledge and skills.

Classroom Feasibility Testing

In preparation for the three-state field test, feasibility tests were conducted in the classrooms of a convenience sample of two teachers, one for each topic, in order to detect logistical challenges associated with delivering assessments and collecting data via the Internet. The Ecosystems classroom feasibility testing took place in four sixth-grade classes with a total of 105 students. The Force & Motion classroom feasibility testing took place in two eighth-grade classes with a total of 33 students. The classroom feasibility test also served as a pilot of the research instruments and data collection procedures. Using a classroom observation form described later researchers recorded the overall conditions of the implementation, student engagement in the tasks, teacher role, and individual and group activity structures, as well as any problems related to the science content or technology use. In addition, twenty-eight students from the classes participated in cognitive labs to verify that the tasks and items were eliciting the intended science content and inquiry skills. Researchers also checked that the database systems were accurately recording the assessment data for each student and that reports generated for students and teachers were accurate.

During feasibility testing, all students completed the embedded assessments in <20 minutes and the benchmark in <45 minutes. This indicated that the assessments were appropriate for the class periods in schools, which are typically about 50 minutes long.

Some technical problems were detected, which led to changes in how the assessments were transmitted to the classroom. For example, when students begin an assessment, a series of files is stored on the student’s computer and is used to present the assessment tasks. When network bottlenecks slowed download rates, the total length of time required sometimes exceeded the limits set for this process. Limits were therefore adjusted to accommodate slower networks and files were modified to reduce their sizes.

Data from student responses to the assessments, the classroom observations, and input from the teachers informed revisions to the simulation-based embedded assessments, the reflection activities, unit benchmark assessments. Revisions were also made as needed to the classroom observation forms and teacher interview protocols.

Methods

The goals of the large-scale field test were to establish the psychometric quality of the SimScientists assessments, the feasibility of implementing them in the classroom, differential student performance (in particular, for ELL and SWD students), their utility for teachers, and to propose models for integrating simulation-based assessments into state assessment systems. Two suites of SimScientists curriculum-embedded and benchmark assessments were tested; one for Ecosystems and one for Force & Motion.

The field test sought to answer four research questions:

1. Are the simulation-based assessments of sufficient technical quality (reliability and validity) to warrant inclusion in balanced state assessment systems?
2. Do they work well for English language learners and students with disabilities?
3. Is it feasible to implement these assessments in a diverse array of classrooms?
4. Do teachers find them useful in monitoring and adjusting instruction for their students?

Participants

Three states (North Carolina, Nevada, and Utah) of the six states on the Design Panel volunteered to participate in the field test. The Design Panel members, from the education agencies of each state, assisted in the recruitment of teachers by distributing project information and applications. All teachers who chose to take part in this research project were accepted, subject to their school's ability to implement the web-based assessment materials. Teachers were required to attend the project's professional development sessions and complete specific data collection requirements in order to receive a stipend.

Participants included 55 teachers and 5,867 students in 28 districts and 39 schools. Schools included large urban settings, small rural schools, charter schools, and a juvenile detention facility. A total of 3,529 students took part in the test of the Ecosystems assessments and 1,936 students tested Force & Motion. One state did not implement the Force & Motion assessments because the standards targeted by the assessments did not align sufficiently with their middle school standards.

Students were approximately evenly divided between males and females. Of the 5,660 students for whom we have complete data, approximately 12% were identified by the school as students with disabilities in one of two ways. Roughly 11% had Individualized Education Programs (IEPs); <1% had accommodation plans required by Section 504 of the Rehabilitation Act of 1973 (504 plans). Approximately 6% of the students were classified as English language learners by their schools. Approximately 34% were eligible for free or reduced-price lunch. Ethnicities represented included Caucasian (66%), Hispanic (13%), African-American (11%), Asian (4%); the remaining 6% were identified as multiracial, native American, Pacific Islanders, or unknown ethnicity.

Data Collection

Instruments. The following sections describe the data collection instruments and procedures, most of which had been used in our prior research. The descriptions of the instruments are organized by pairs of research questions.

- (1) Are the simulation-based assessments of sufficient technical quality (reliability and validity) to warrant inclusion in balanced state assessment systems?
- (2) Do they work well for English language learners and students with disabilities?

SimScientists Simulation Assessment Data. To examine the psychometric quality of the simulation-based assessments, the LMS captured data as students completed the assessments and stored the data in a secure database. These data include observable events, such as answers to questions, inputs to simulations, the full text of constructed responses, and arrows drawn. Each observable event was coded by content or inquiry target. These data were then analyzed in real time using algorithms and rubrics created by the design team, and classified as evidence of misconceptions, types of errors, or a correct understanding and coded to the appropriate diagnostic variable. Both the observable events and the diagnostic variables were sent to the database and used to generate reports.

The curriculum-embedded formative assessments also captured and analyzed the type and amount of help (feedback and coaching) that students needed to complete assessment tasks. From these data, the LMS parsed students into three groups, (A) those who needed no feedback or only minimal feedback that indicated an error without providing any coaching, (B) those who typically needed coaching that describes the scientific principles to be applied, and (C) those who often needed worked examples before they could respond correctly. These categories were intended to assist teachers in making subsequent decisions about differentiating additional instruction.

During benchmark assessment use, student response data were collected from the simulations in the form of a dichotomous or polytomous score for each measured response or action and coded to a content or inquiry target, such that overall ability could be calculated for each of these two dimensions. Similarly, in the posttest, students' dichotomous scores on the multiple-choice items were recorded and coded to the relevant science content or inquiry targets.

The end-of-unit benchmark assessments combined items scored electronically along with teacher scores of written responses. Teachers received online scorer training before blind-scoring their students' written responses (no names were associated with student responses). The LMS combined these scores in a single record for each student. The assessments used Bayes' Nets to estimate student proficiency on content and inquiry targets. On the summative benchmark assessments, the LMS classified students as Advanced, Proficient, Basic, or Below Basic by model level and by content and inquiry targets.

Posttests. External measures of science knowledge and inquiry practices were constructed for both middle school topics. Thirty items were selected from the AAAS Project 2061 item database (<http://assessment.aaas.org/>). Approximately one-fourth of the items targeted the inquiry practice *control of variables*. The remaining items were distributed evenly among the three system levels (components, interactions, emergent behavior) for each topic. Using AAAS data about the percentage of students who correctly answered the questions (also available on the website), the posttests were constructed to have a range of item difficulty. Items were sequenced to begin the test with a few easier items, followed by a distribution of more difficult items and inquiry items throughout the remainder of the posttest.

Student Demographic Data. Demographic data available varied by state and included gender, ethnicity, free, and reduced-price lunch, English proficiency levels of ELL students, and disability categories of students with IEPs or 504 plans. Analyses reported here used students' ELL status, and whether students had IEPs or 504 plans.

- (3) Is it feasible to implement these assessments in a diverse array of classrooms?
- (4) Do teachers find them useful in monitoring and adjusting instruction for their students?

Teacher Pre-Surveys. Prior to the professional development sessions, teachers completed surveys in which they indicated the amount of time they planned to spend on each of the content and inquiry targets, the instructional approaches they planned to use, their plans for accommodations, and predictions about the feasibility of the assessment system. These data were used during teacher professional development meetings to help teachers determine when to insert the simulation-based assessments and reflection activities during the unit.

Teacher Online Surveys. Teachers completed online surveys after each use of the assessments. Teachers indicated the extent of prior instruction, any technical difficulties, and summarized observations of student use and reactions. These data included the number of class periods spent on each of the content and inquiry targets during the unit, the types of instructional approaches used (such as hands-on activities, lectures, or other computer-based instruction), how teachers acted on data from the embedded assessment reports, with particular focus on how the reflection activities were implemented, the reports' utility for making instructional decisions, and what accommodations were used during instruction, if any, for students who needed them.

Case Studies. CRESST conducted case studies in a convenience sample of five schools across the three states. The case studies used data from 56 classroom observations and eight teacher interviews (Herman, Dai, Htut, Martinez, & Rivera, 2010). These instruments are described below.

CLASSROOM OBSERVATIONS. The classroom observations were intended to monitor fidelity of implementation of the assessments. In 5-minute time samples, trained observers documented student engagement in the tasks (actively engaged, passively watching, or off task), teacher role, and individual and group activity structures, as well as any problems related to the science content or technology use and the overall conditions of the implementation. Data were collected about how the teacher assisted individual students or addressed the entire class. Data collected about overall conditions included whether students worked alone or in pairs, whether the setting was a classroom or computer lab, and the number of computers in use.

TEACHER INTERVIEWS. Interviews were conducted following the classroom observations probing in more depth about teachers' evaluations of the assessment implementation. Teachers were asked questions about the feasibility of implementing the assessments and the utility of the reports for making instructional decisions and for gauging their students' achievement on the content and inquiry targets.

Intervention

Materials. As described in greater detail earlier, the Ecosystem assessment suite consisted of two simulation-based embedded assessments and a unit benchmark assessment. The Force & Motion suite consisted of three embedded assessments and a benchmark assessment. Each embedded assessment required one period for the simulation-based activities and one follow-up period in the classroom for the reflection activity. Each simulation-based benchmark assessment required one period, as did the 30-item posttest. Thus, the Ecosystem suite took place in seven class periods, not including the teachers' regular instruction on the topic, while the Force & Motion suite required nine class periods. All online assessments supported screen magnification and audio accommodations, and segmentation for stopping and re-entering the assessments.

Professional Development. Teachers in the field test participated in 12 hours of professional development (PD) conducted face-to-face in their nearby schools. Teachers were oriented to the research needs of the project and its principles and goals. Teachers were provided with alignments of the assessments to their state science standards along with tools to map the alignment of the assessments to the learning goals of their classrooms, allowing them to decide the appropriate times for embedding the assessments in their curriculum units. Teachers worked through each of the embedded and benchmark assessments and participated in a reflection activity modeled by the PD leader. They learned how to use the LMS to register their classes for the simulation activities, assign accommodations to ELLs and SWDs based on their state and district testing guidelines, interpret the embedded assessment progress reports, group students for reflection activities, score constructed responses from the benchmark assessment, and interpret the benchmark proficiency reports. Following the PD session, teachers completed a questionnaire evaluating it.

Implementation. Teachers were recruited for the field test beginning in Fall 2009. Professional development was conducted in Spring 2010 shortly before data collection commenced in classrooms. Teachers notified the project technology coordinator of their schedules for implementing the embedded and benchmark assessments and began the process of registering their classes and assigning accommodations. Project staff worked with teachers and technology coordinators to ensure that the assessments would run on schools' networks; a "Help Desk" provided assistance as needed during implementation.

At the appropriate time in the curriculum, the teacher assigned the embedded assessments to students, who worked through the embedded assessment either in a computer lab or in the classroom using laptops. For the embedded assessments, students were able to work individually on the computers, in pairs on one computer, or in pairs on individual computers. When students had completed each embedded assessment, the teacher used the LMS to review the progress reports for the class and for individual students. These progress reports suggested appropriate team assignments for students based on the amount of help needed, and whether the help was for tasks aligned primarily to content or inquiry targets. For example, in the Ecosystems reflection activity that follows the Food Web embedded assessment, students who had difficulty identifying the producers and consumers in the food web were given more practice in that task while identifying consumers in the reflection activity (in a new ecosystem: Galapagos, savannah, or tundra) and would then combine their findings with other teams to produce a food web for one of the three new ecosystems. These groups then presented a description of the organisms, their roles, and a food web for their ecosystem to the entire class. Thus, reflection activities provided additional differentiated instruction and opportunities for scientific discourse and collaboration. Teachers completed online surveys after each pair of components in the suite, comprised of an embedded assessment and a reflection activity.

Teachers engaged in their usual instruction before assigning students to the next embedded assessment and reflection activity. When students had completed all embedded assessments, the teacher assigned the summative benchmark assessment and the on-line multiple-choice posttest. These were completed by students individually. Teachers completed a final survey after completing all components of the suite.

Case studies were conducted by CRESST using a convenience sample of eight teachers teaching either Ecosystems or Force & Motion in five schools across the three states: two schools and three teachers in Utah, two schools and three teachers in Nevada, and one school and two teachers in North Carolina. Evaluators received training in conducting the

observations and the use of the teacher interview protocol. CRESST conducted 56 classroom observations representing the full range of project activities and conducted interviews with the eight teachers (Herman et al., 2010).

Analyses

Feasibility and utility were examined by teacher surveys, computer logs, and the case studies conducted by the external evaluator, CRESST. Descriptive statistics summarized assessment completion rates from computer logs, teacher responses about the quality and utility of the assessments on the surveys, frequencies of categories of observed teacher and student activities and engagement, and common themes in teacher interviews.

Technical quality of the assessment system was examined primarily through analyses of student responses to the assessments. To determine whether the categorizations of students from the embedded assessments were reasonable, the assignments of students to the different groups, A, B, and C, in the embedded assessments were analyzed to see if the groups differed in their performances on the benchmark assessments. To judge the performance of the assessment items and the overall reliability of the assessment system, a multi-dimensional partial credit Item Response Model (IRT) was fitted to the benchmark response data. In cases where multiple items shared a single stimulus, items were bundled for analysis because the items could not be considered as truly independent measures. Bundles were treated as polytomous items, with a minimum score of 0 and a maximum score equal to the total number of items in the bundle. For example, there are eight possible response patterns for a bundle of three items, (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), and (1,1,1). Validity of the assessments was established by triangulating psychometric analyses with judgments by AAAS of the alignment of the assessments to national and state standards, cognitive labs conducted with a sample of 28 students from the classroom feasibility testing to provide evidence of construct validity, that is, that the simulations elicited expected thinking about science content and related inquiry practices, and analyses of performance on the assessments (Quellmalz et al., 2005). Evidence of concurrent validity was obtained by comparing performance on the simulation benchmark assessments to performance on the 30-item posttest of multiple-choice items drawn from the AAAS bank of calibrated items. Estimates of student ability from the benchmark assessments were compared with the independent posttest results to examine correlations. Discriminative validity was inspected by examining the extent to which the simulation-based assessments distinguished between science content and inquiry constructs more effectively than did the conventional posttest.

Results

Feasibility and Utility

Feasibility and utility were primarily documented by the CRESST evaluation (Herman et al., 2010). The case studies found that teachers were able to implement the computer-based assessments and that students were highly engaged in the SimScientists assessments and able to complete them successfully. The computer logs confirmed that students completed the assessments within the allotted class period. The observations confirmed that teachers were implementing the computer-based assessments and reflection activities as intended.

Responses to the surveys and interviews indicated that, overall, both teachers and students responded favorably to the SimScientists embedded formative and end-of-unit benchmark assessments. Teachers rated nearly all of the questions on their surveys 3 or higher on a 4-point scale. Observations provided evidence that students were active and engaged during

the assessments and that teachers gave positive feedback when interviewed. Teachers collectively agreed that the simulation-based assessments had greater benefits than traditional paper-and-pencil tests because of the simulations' instant feedback, interactions, and visuals. Logistically, most teachers stated that they needed computers to be more easily accessible in order to implement the computer assessment several times in the year. Observations and interviews, both with teachers and state Design Panel members, suggested that teachers and students were highly satisfied with SimScientists and able to implement the assessments effectively.

Technical Quality

Weighted Mean Square Fit statistics from the IRT analysis of the student responses on the Ecosystems benchmark assessment were between .8 and 1.2 for all except one of the 45 items, fitting the measurement model and contributing information relevant to the overall measure of science content and inquiry practices. The reliability was .76 for the ecosystems benchmark assessment, which is considered acceptable (George & Mallery, 2003), particularly for an assessment that was a mixture of selected response, interactions with the simulations, and short written responses scored by the teachers. Similarly, for the Force & Motion benchmark assessment all except 1 of the 41 items fitted the measurement model, which indicated that almost all of items were contributing information relevant to the overall measure. The reliability for the Force & Motion benchmark assessment was .73, which is acceptable. Empirical data from the IRT analyses are available in the Supporting Information.

Evidence of validity of the simulation-based assessments came from several sources. The review by content experts at AAAS confirmed that the assessment tasks were aligned to important content and inquiry targets as defined by the national and state science standards. Cognitive labs contributed evidence of construct validity. The analysis of think-aloud sessions with 28 students during usability testing and the classroom feasibility tests consisted of judgments by raters that the students were applying the intended content and inquiry skills. An average of 84% of the items were judged to elicit the targeted knowledge and inquiry practices as students worked through the tasks. Items not eliciting responses about the targeted content and inquiry were revised prior to the field test.

A one-way ANOVA was used to test for differences among the three classifications of students (groups A, B and C) in their performance on the benchmark assessments. Data from the first Ecosystems embedded assessment showed that performances on the Ecosystems benchmark differed significantly across the three classification groups on both science content, $F(2, 2729) = 338.30, p = .000$ and on inquiry practices $F(2, 2729) = 23.21, p = .000$. Similarly, for the second Ecosystems embedded assessment performances on the Ecosystems benchmark differed significantly across the three classification groups on both science content, $F(2, 2737) = 153.36, p = .000$ and on inquiry practices $F(2, 2737) = 29.85, p = .000$. Likewise, data from the first Force and Motion embedded assessment showed that performances on the Force and Motion benchmark differed significantly across the three classification groups on both science content, $F(2, 1341) = 64.92, p = .000$ and on inquiry practices $F(2, 1341) = 100.99, p = .000$. Similarly, for the second Force and Motion embedded assessment, performances on the Force and Motion benchmark differed significantly across the three classification groups on both science content, $F(2, 1262) = 97.19, p = .000$ and on inquiry practices $F(2, 1262) = 83.70, p = .000$. The pattern was repeated for the third Force and Motion embedded assessment performances that differed significantly for the three classification groups on science content, $F(2, 1281) = 72.04, p = .000$ and on inquiry practices $F(2,$

1281) = 83.98, $p = .000$. Overall this shows that classifications of students in the embedded assessments into three groups was valid in that the classifications were reflected in significant differences in performance on the benchmark test. ANOVA tables are available in the Supporting Information.

Further concurrent validity evidence came from the correlation of the student performance on the science content and inquiry measures from the benchmark assessment with their performances on the independent posttest. All four of the correlations were statistically significant, although they were moderate (from .57 to .64) indicating that the benchmark and posttest assessments measured similar science content and inquiry practices, but that the measures were not exactly the same. This was expected as the simulation-based assessments were designed to measure content knowledge and skills that cannot be assessed fully with conventional items. In particular, the correlations for inquiry were lower than the correlations for content, supporting this interpretation.

Discriminative validity was established by analyses finding that the benchmark assessment distinguished student performance on inquiry practices more effectively than the posttest. The correlation of the content and inquiry dimensions on the posttest for Ecosystems (.85) and Force & Motion (.92) were higher than those for the benchmark assessments (.70 and .80, respectively). This indicates that the discrimination between the measures of content and inquiry is greater in the simulation-based benchmark assessment than on the traditional items of the posttest. Additional detail on these results is available in the Supporting Information.

Performance of English Language Learners and Students With Disabilities. Overall, students performed better on the benchmark assessments than on the posttest, and performance gaps between both ELLs and SWDs compared to other students were reduced on the benchmark. To determine the effect of the simulation-based assessments on ELLs and SWDs, their performances on the benchmark assessments were compared to performance on the posttest of conventional items. Table 1 compares performance gaps of ELLs and SWDs to a reference group of all students who are neither English language learners nor students with disabilities. Although the average performances of ELLs and SWDs on the SimScientists benchmark is lower than that of the reference group, the gaps between the focal groups and the reference group is comparatively smaller than for the post test. This evidence provides some support for the claim that the multiple representations in the simulations and active manipulations may have provided alternative means, other than written text, for ELLs and SWDs to understand the assessment tasks and questions and to respond.

The differences in the performance gaps were even more marked in the measurement of the science inquiry skills, as shown in Table 2. There were much larger performance gaps on the inquiry skills on the posttests than there were on the benchmark assessments. This evidence suggests that the benchmark assessments allowed ELLs and SWD to demonstrate their

Table 1
Gaps in total performance between English learners or students with disabilities and the general population

Group	Ecosystems Posttest	Force & Motion Posttest	Ecosystems Benchmark	Force & Motion Benchmark
English learners	24.0% ($n = 123$)	27.4% ($n = 50$)	10.6% ($n = 126$)	13.6% ($n = 50$)
Students with disabilities	20.2% ($n = 183$)	15.7% ($n = 153$)	8.4% ($n = 189$)	7.0% ($n = 153$)

Table 2

Gaps in inquiry skills performance between English learners or students with disabilities and the general population

Group	Ecosystems Posttest	Force & Motion Posttest	Ecosystems Benchmark	Force & Motion Benchmark
English learners	25.6% (<i>n</i> = 123)	35.1% (<i>n</i> = 50)	6.6% (<i>n</i> = 126)	10.9% (<i>n</i> = 50)
Students with disabilities	25.5% (<i>n</i> = 183)	20.3% (<i>n</i> = 153)	5.6% (<i>n</i> = 189)	6.2% (<i>n</i> = 153)

inquiry skills more clearly in the simulation-based benchmark assessments than they were in the multiple-choice item posttests. The benefits of simulations for these groups warrant further investigation.

Models for Integrating Simulation-Based Science Assessments Into a Balanced State Science Assessment System

The primary goals of the study were to determine the suitability of simulation-based assessments for a state science assessment system and to describe models for incorporating them. The six-state Design Panel reviewed the field test findings supporting the technical quality, feasibility, and utility and judged that the SimScientists simulation-based assessments could serve as credible components of a state science assessment system. Interviews of the state representatives by CRESST documented positive feedback overall. The state representatives reported that the SimScientists assessments worked well, and that teachers were willing to participate. The state representatives shared feedback from teachers that they were impressed with the software and activities and would welcome the opportunity to participate again. Given the teachers' reactions and the nature of the assessments and associated reflection activities, the state representatives were interested in knowing plans and topics for future development and likely topics to be developed. They also encouraged development and implementation in subject areas beyond science, such as mathematics.

Representatives on the Design Panel collaborated with WestEd to formulate two models for states' use of simulation-based science assessments. The models aimed to describe how the simulation-based assessments could become part of a balanced state assessment system by using them at the classroom, district, and state levels with common designs that would make them mutually reinforcing (Pellegrino et al., 2001; Quellmalz & Moody, 2004). The two models created combinations of simulation-based science assessments that would be *coherent* with each other, *comprehensive* in coverage of state science standards, and provide *continuity* of assessments through multiple forms and occasions.

The two models proposed include (1) using the unit benchmark assessment proficiency data to augment state reports and (2) using a sample of simulation-based shorter, signature tasks parallel to those in the benchmarks administered as part of state or district tests. Figure 4 presents a sample report that could be generated in the "Side-by-Side" model in which data at the state, district, and classroom levels are mutually aligned and complementary. District and classroom assessments can provide increasingly rich sources of information, allowing a fine-grained and more differentiated profile of a classroom, school, or district that includes aggregate information about students at each level of the system. In this "Side by Side" model, the unit benchmark assessments can function as multiple measures administered after science units during the school year, providing a continuity of in-depth, topic-specific "interim" or "through-course" measures that are directly linked in time and substance to units on science systems such as climate or earth's structure.

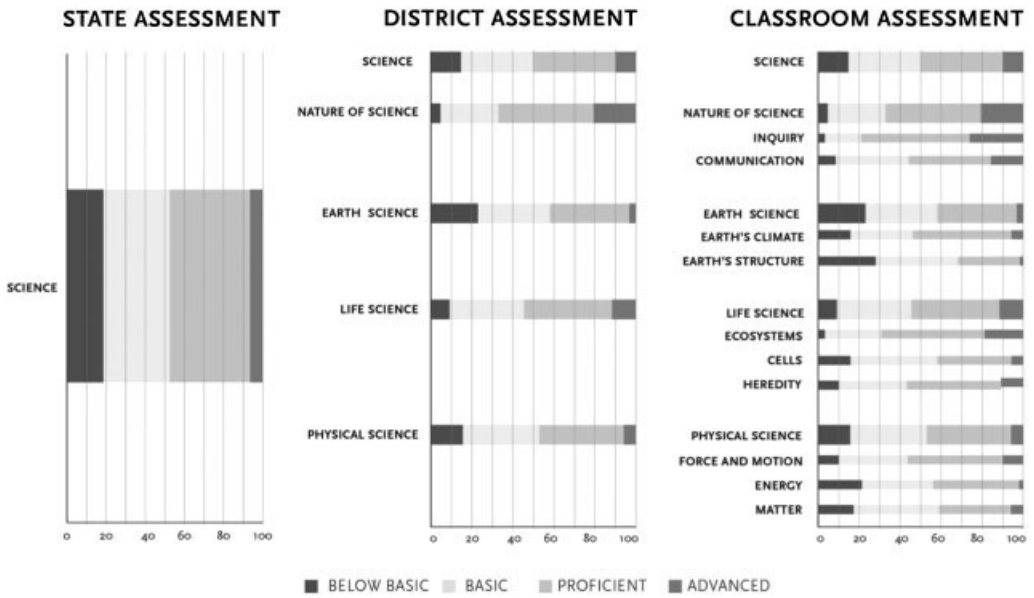


Figure 4. Side-by-Side Model, showing how data reported from unit benchmark assessments can augment information from district and state science reports.

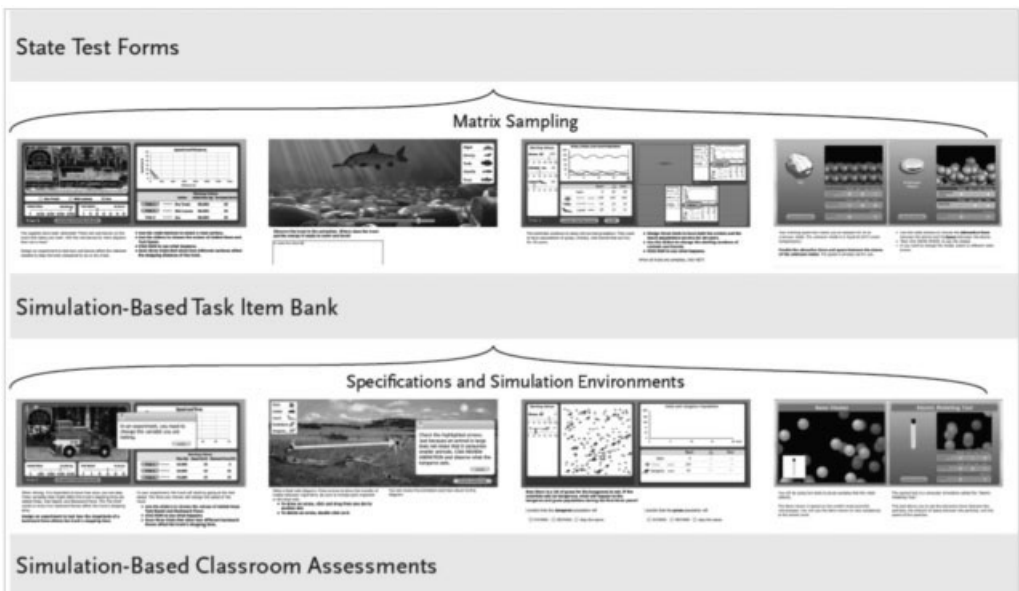


Figure 5. Signature Task model, showing how parallel tasks can be developed for state and classroom assessments.

Figure 5 portrays the “Signature Task” model in which states and districts draw upon the specifications and rich simulation environments developed for the classroom-level unit benchmark assessments to create a new, parallel set of key, or signature, tasks such as drawing a food web, or conducting a predator-prey investigation. The classroom-level simulation-based tasks might be set in a mountain lake ecosystem, while parallel tasks developed for state or district tests would be set in different ecosystems such as the grasslands or tundra. These signature tasks could be administered in a matrix sampling design during the state or district testing to collect data on inquiry practices and integrated knowledge not fully measured by traditional item formats on the state test.

For example, the first task in each row shows a signature task for inquiry into the effect of forces on objects. On the state test, the object is a train. On the classroom assessment, the object is a fire truck. The masses, forces, and results of the investigations vary between the parallel tasks, but the simulation interface and the inquiry task structure are otherwise identical.

This model assures coherence of assessment task types in the different levels of the assessment system. The two models can provide a template for states to begin moving closer to the goal of a system for state science assessment that provides meaningful information drawn from a system of nested assessments collected across levels of the educational system.

As a culminating activity of the Enhanced Assessment Grant, the project produced a Policy Brief, which summarized the study results and presented the two integration models. The Policy Brief was disseminated to policymakers in the participating states and at national conferences, and is available on the SimScientists website (www.simsScientists.org).

Discussion

For many science assessments, particularly those used to document complex learning in simulations, the intended learning goals are tested by static, conventional items not well aligned with specified outcomes (Quellmalz, Timms, & Schneider, 2009). The major contribution of this study is the power to assess important science knowledge and inquiry skills with high psychometric quality and at large scale. This was achieved by a synthesis of the model-based learning framework and evidence-centered design. We conceptualized the content in terms of complex systems (components, interactions, and emergent behaviors) and inquiry practices in terms of practices engaged in by scientists (Buckley, in press). We used a generalizable design framework, which we have used in other content areas, and have tested its utility for organizing the structure and content for both assessment and instruction. The SimScientists assessments employed the rigor of evidence-centered assessment design to closely link the knowledge and processes to be assessed, to the evidence of learning and task features focused on eliciting the evidence. Further, the capability of these assessment designs to discriminate performance on content and inquiry will provide educators with evidence of the development of inquiry practices currently missing in most assessments (Quellmalz et al., 2005). Furthermore, multilevel assessments linked by common specifications will strengthen the coherence and vertical articulation among assessments (Quellmalz, Timms, & Silbergliitt, 2011). The re-use of simulation environments and design specifications for assessments at each level (classroom, district, state) will ensure this coherence while reducing development cost.

This study addressed a number of recurring criticisms about the suitability of technology-enhanced assessments and complex assessment tasks for widespread classroom use or for accountability. Critics worry that the variability of school technical infrastructures and levels

of teacher and student comfort with technology will thwart effective implementation of technology-based assessments. In this study the simulation-based science assessments were successfully implemented across diverse settings, teachers, and students. Critics also contend that English Language Learners and students with disabilities may be overwhelmed by complex, dynamic simulations. In this study, evidence suggests that these populations of students were better able to show what they know and can do on the simulation-based benchmark assessments than on the posttest composed of static, conventional items.

The curriculum-embedded simulation assessments did serve formative purposes as evidenced by the implementation evaluation. Teachers indicated that the embedded assessment progress reports prompted adjustment of subsequent instruction during the unit. Both teachers and students commented on the value of the immediate, individualized feedback, and coaching. The coaching provided scaffolding in the form of additional instruction that strengthens the learning benefit of the curriculum-embedded assessments. The embedded simulation-based assessments can thus serve as a powerful resource in a teacher's formative assessment tool kit.

Critics suggest that complex, technology-enhanced assessments will not have sufficient technical quality to be appropriate for accountability purposes. This study provided evidence that the simulation-based unit benchmark assessments had high technical quality (reliability and validity) and could, therefore, be used in district or statewide reports of achievement on science standards. Moreover, the fact that science assessment coordinators from six states contributed to the development of two models for integrating formative and summative simulations into their state science test systems suggests that other states may also well be ready to consider incorporating science simulations in the next generation of assessments into their state science assessment systems.

Overall, the study provided strong evidence for the claim that the SimScientists simulation-based assessments could be scaled up to serve as credible components of a state science assessment system. Although the quality of the assessments was demonstrated for two topics, the data support the promise of simulation-based assessments across the curriculum. The systematic and principled design process coupled with reusable design templates further support efforts to increase the use of dynamic science assessments in classroom, district, and state assessment programs. Therefore, this study should be of interest to the state assessment consortia that are aiming for such balanced, multilevel assessment systems (Darling-Hammond & Pecheone, 2010).

Limitations

The authors acknowledge a number of limitations. The EAG project provided an opportunity to field test simulation-based assessment suites in classrooms across three states. The states and teachers were self-selected volunteers. The state science assessment programs were committed to examining technology-enhanced formative and summative assessment possibilities. The participating schools had the necessary technology infrastructures. Importantly, the assessment results were not intended for high stakes accountability. In addition, teachers received stipends for their participation. We realize that teacher and student perceptions of the worth of the assessments might be affected by support from policymakers and by the stakes attached to the use of the results.

The implementation evaluation conducted by CRESST used a convenience sample for the case studies, driven by travel and budget constraints, so the 8 case study teachers were not necessarily representative of the 55 field test teachers. However, the computer logs of student performance on the simulation-based assessments and the online surveys completed

by all teachers provided data on the feasibility and utility of the assessments for the entire field test sample.

Another limitation of the study is that the simulation-based assessments addressed one science unit within each of two grade levels. If simulation-based assessments suites were available for multiple units during a yearlong curriculum, schools would need to orchestrate access to computer labs over the course of the year. Access to computers will remain an issue in the short term, although access is increasing, especially as state assessment consortia prepare to administer state tests online.

While the benefits of the formative embedded assessments were recognized and acknowledged by the teachers in this study, research is needed to document the effectiveness of the embedded assessments for supporting learning. That research is currently underway in the Calipers II project, which is conducting small randomized-control trials in classrooms.

Data on the use of the accommodations and effects of the simulations for English language learners was limited. Future studies would need to ensure larger samples of populations that could benefit from such accommodations.

Conclusion

This study is one of the few to provide research-based evidence that systematically developed simulation-based science assessments used for formative and summative purposes can achieve high technical quality, be broadly implemented, and have strong instructional utility. Moreover, findings are very promising for the potential benefits of simulations for assessing important science learning and inquiry practices for all students. Support by six states for the value of science simulations for assessing important science standards bodes well for further use of simulations to test the forthcoming Next Generation Science Standards. The study provides evidence to support the recommendation that innovative technology-enhanced assessments can be credible components of multilevel, balanced state science assessment systems.

This article is based upon work supported by the US Department of Education (Grant No. 09-2713-126) and the National Science Foundation (Grant No. 0733345) Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the US Department of Education or the National Science Foundation. Additional publications from this study can be found at <http://simscientists.org>.

References

- American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for psychological testing*. Washington, DC: American Educational Research Association.
- Bangert-Downs, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61(2), 213–238.
- Bell, P., & Linn, M. C. (2000). Beliefs about science: How does science instruction contribute? In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Betrancourt, M. (2005). The animation and interactivity principles. In R. E. Mayer (Ed.), *Handbook on multimedia learning*. New York: Cambridge University Press.

- Black, P. (1998). *Testing: Friend or foe? Theory and practice of assessment and testing*. New York: Falmer Press.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., & Soloway, E. (1996). Learning with peers: From small group cooperation to collaborative communities. *Educational Researcher*, 25(8), 37–40.
- Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy. *Remedial and Special Education*, 25(3), 141–152.
- 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. (2000). Interactive multimedia and model-based learning in biology. *International Journal of Science Education*, 22(9), 895–935.
- Buckley, B. C. (in press) Supporting and assessing complex biology learning with computer-based simulations and representations. In D. Treagust & C.-Y. Tsui (Eds.), *Multiple representations in biological education*. New York: Springer.
- Buckley, B. C., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking inside the black box: Assessing model-based learning and inquiry in biological. *International Journal of Learning Technologies*, 5(2):166–190.
- Buckley, B. C., Gobert, J., Kindfield, A. C. H., Horwitz, P., Tinker, B., Gerlits, B., . . . Willett, J. (2004). Model-based teaching and learning with hypermodels: What do they learn? How do they learn? How do we know? *Journal of Science, Education and Technology*, 13(1), 23–41.
- Burling, K., Beck, R., Jude, J., Murray, E., Harms, M., & Dolan, B. (2006). Constructing innovative computer-administered tasks and items according to universal design: Illustrative examples with pilot data. Paper presented at the Annual Meeting of the National Council on Measurement in Education.
- Case, B. J. (2008). Accommodations to improve instruction, learning, and assessment. In M. Karchmer R. Johns & M. Ross (Eds.), *Testing deaf students in an age of accountability*. Washington, DC: Gallaudet Research Institute.
- Case, B. J., Brooks, T., Wang, S., & Young, M. (2005). *Administration mode comparability study*. San Antonio, TX: Harcourt Assessment, Inc.
- CAST. (2008). *Universal design for learning guidelines version 1.0*. Wakefield, MA: Author.
- Chiu, C. W. T., & Pearson, P. D. (1999). Synthesizing the effects of test accommodations for special education and limited English proficiency students. Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.
- Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *Handbook of creativity: Assessment, theory and research* (pp. 341–381). New York: Plenum Press.
- College Board. (2009). *Science: College Boards standards for college success*. Retrieved from <http://professionals.collegeboard.com/profdownload/cbscs-science-standards-2009.pdf>.
- Collins, A., Brown, J. S., & Newman, S. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In: L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Rober Glaser* (pp. 453–494). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Covington, M. V. (1999). Caring about learning: The nature and nurturing of subject-matter appreciation. *Educational Psychologist*, 34(2), 127–136.
- Darling-Hammond, L. (2010). *Performance counts: Assessment systems that support high-quality learning*. Washington, DC: CCSSO.
- Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Presented at the National Conference on Next Generation K–12 Assessment Systems, Center for K-12 Assessment & Performance Management with the Education Commission of the States (ECS) and the Council of Great City Schools (CGCS), Washington, DC
- Dassa, C., Vazquez-Abad, J., & Ajar, D. (1993). Formative assessment in a classroom setting: From practice to computer innovations. *Alberta Journal of Educational Research*, 39(1), 111–125.

- Donovan, M. S., & Bransford, J. D. (2005). *How students learn: Science in the classroom*. Washington, DC: The National Academies Press.
- Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, 38, 39–72.
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.
- Edelson, D., & Reiser, B. J. (2006). Making authentic practices accessible to learners: Design challenges and strategies. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 335–354). New York: Cambridge University Press.
- Forsyth, B., & Lessler, J. T. (2004). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*. Hoboken, NJ, USA: Wiley.
- Geier, R., Blumenfeld, P., Marx, R., Krajcik, J., Fishman, B., & Soloway, E. (2008). Standardized test outcomes of urban students participating in standards and project based science curricula. *Journal of Research in Science Teaching*, 45(8), 922–939.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update (4th ed.) Boston: Allyn & Bacon.
- Gerard, L. F., Spitulnik, M., & Linn, M. C. (2010). Teacher use of evidence to customize inquiry science instruction. *Journal of Research in Science Teaching*, 47(9), 1037–1063.
- Gobert, J. (2000). A typology of models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education*, 22(9), 937–977.
- Gobert, J. D., & Clement, J. J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching*, 36(1), 39–53.
- Goldstone, R. L. (2006). The complex systems see-change in education. *The Journal of Learning Sciences*, 15, 35–43.
- Goldstone, R. L., & Wilensky, U. (2008). Promoting transfer through complex systems principles. *Journal of the Learning Sciences*, 17, 465–516.
- Harns, M., Burling, K., Hanna, E., & Dolan, B. (2006). *Constructing innovative computer-administered tasks and items according to universal design: Establishing guidelines for test developers*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.
- Heiser, J. (2004). *External representations as insights to cognition: Production and comprehension of text and diagrams in instructions*. Unpublished doctoral dissertation. Stanford University.
- Herman, J., Dai, Y., Htut, A. M., Martinez, M., & Rivera, N. (2010). *CRESST evaluation report: Evaluation of the Enhanced Assessment Grants (EAGs)*. Los Angeles: CRESST.
- Herman, J. L., Osmundson, E., Ayalya, C., Schneider, S., & Timms, M. (2005). *The nature and impact of teachers formative assessment practices*. Paper presented at the American Educational Research Association.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141–158.
- Horwitz, P., Gobert, J. D., Buckley, B. C., & O'Dwyer, L. M. (2010). Learning genetics with dragons: From computer-based manipulatives to hypermodels. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences*. New York: Springer.
- Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, 6(3), 41–49.
- King, K. (2011). *Balanced, multilevel science assessment systems*. Presented at the National Conference on Student Assessment. Orlando, FL.
- Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting learning by design into practice. *The Journal of the Learning Sciences*, 12(4), 495–547.

Kopriva, R. (2008). *Improving testing for English language learners*. New York: Routledge—Taylor & Francis Group.

Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., & Fredricks, J. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *Journal of the Learning Sciences* 7(3&4), 313–350.

Lee, O. (2002). Science inquiry for elementary students from diverse backgrounds. In W. G. Secada (Ed.), *Review of research in education* (Vol. 26, pp. 23–69). Washington, DC: American Educational Research Association.

Lehrer, R., & Schauble, L. (2002). Symbolic communication in mathematics and science: Co-constituting inscription and thought. In E. D. A. J. Byrnes (Ed.), *Language, literacy, and cognitive development. The development and consequences of symbolic communication* (pp. 167–192). Mahwah, NJ: Lawrence Erlbaum Associates.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In D. K. S. Carver (Ed.), *Cognition and instruction: 25 years of progress* (pp. 39–74). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M. C., Bell, B., & Davis, E. A. (2004). *Internet environments for science education*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Maehr, M. L., & Midgley, C. (1996). *Transforming school cultures*. Boulder, CO: Westview Press.

Martin, J., & VanLehn, K. (1995). Student assessment using bayesian nets. *Journal of Human-Computer Studies*, 42(6), 575–591.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, 41(10), 1063–1080.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology*, 84(4), 444–452.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction*, 22(2):219–290.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 12–23.

Minnesota Department of Education. (2010). *Draft test specifications for science*. Available at: http://education.state.mn.us/MDE/Accountability_Programs/.Assessment_and_Testing/Assessments/MCA/TestSpecs/index.html.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439–483.

Mislevy, R., & Haertel, G. D. (2007). Implications of evidence-centered designs for educational testing. *Educational Measurement: Issues and Practices*, 25(4), 6–20.

National Assessment Governing Board (NAGB). (2008). *Science framework for the 2009 national assessment of educational progress*. Washington, DC: Author.

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

National Research Council (NRC). (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge MA: MIT Press.

Nolin, M., & Chandler, K. (1996). Use of cognitive laboratories and recorded interviews in the national household education survey (No. NCES-96-332). Rockville, MD: Westat.

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

- Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing instruction and study to improve student learning (No. NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
- Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.
- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues & Practice*, 28(3), 5–13.
- Phillips, S. E. (1993). Legal implications of high-stakes assessments: What states should know. Oak Brook, IL: North Central Regional Laboratory.
- Quellmalz, E. S., DeBarger, A., Haertel, G., & Kreikemeier, P. (2005). Validities of science inquiry assessments: Final report. Menlo Park, CA: SRI International.
- Quellmalz, E. S., & Haertel, G. (2004). Technology supports for state science assessment systems. Unpublished manuscript, Washington DC.
- Quellmalz, E. S., & Moody, M. (2004). *Models for multi-level state science assessment systems*. Report commissioned by the National Research Council Committee on Test Design for K-12 Science Achievement.
- Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323, 75–79.
- Quellmalz, E., Schank, P., Hinojosa, T., & Padilla, C. (1999). Performance assessment links in science (pals) (No. ERIC Digest Series E-DO-TM-99-04). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (2009). Assessment of student learning in science simulations and games. Washington, D. C.: National Research Council.
- Quellmalz, E., Timms, M., & Silberglitt, M. D. (2011). Integrating simulation-based assessments into state science assessment systems. San Francisco: WestEd.
- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, 13(3), 337–386.
- Rebetz, C., Sangin, M., Betrancourt, M., & Dillenbourg, P. (2004). Effects of collaboration in the context of learning from animations. *Proceedings of EARLI SIG meeting on Comprehension of Text and Graphics: Basic and applied issues*, 9–11, September 2004, Valencia (Spain), 187–192.
- Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3), 273–304.
- Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Khlar (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 263–305). Mahwah, NJ: Erlbaum.
- Rieber, L. P., Tzeng, S., & Tribble, K. (2004). Discovery learning, representation, and explanation within a computer-based simulation. *Computers and Education*, 27(1), 45–58.
- Rivet, A., & Krajcik, J. S. (2004). Achieving standards in urban systemic reform: An example of a sixth grade project-based science curriculum. *Journal of Research in Science Teaching*, 41(7), 669–692.
- Rose, D. H., & Meyer, A. (2000). Universal design for learning. *Journal of Special Education Technology*, 15(1), 67–70.
- Scalise, K., Timms, M., Moorjani, A., Clark, L., Holtermann, K., & Irvin, P. S. (2011). Student learning in science simulations: Design features that promote learning gains. *Journal of Research in Science Teaching*, 48(9), 1050–1078.
- Schwartz, D. L., & Heiser, J. (2006). Spatial representations and imagery in learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*. Cambridge: Cambridge University Press.
- Simon, H. A. (1980). Problem solving and education. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 81–96). Hillsdale, NJ: Erlbaum.

Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn* (pp. 515–565). Washington, DC: The National Academies Press.

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

Stiggins, R. (2006). *Balanced assessment systems: Redefining excellence in assessment*. Educational Testing Service. Portland: Oregon.

Twing, J. S., & Dolan, R. P. (2008). *UD-CBT guidelines*. Retrieved March 5th, 2008 from <http://www.pearsonedmeasurement.com/cast/index.html>.

Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Jordan, R., Gray, S., & Sinha, S. (2011). Understanding complex natural systems by articulating structure-behavior-function models. *Journal of Educational Technology & Society*, 14(1), 66–81.

Vye, N. J., Schwartz, D. L., Bransford, J. D., Barron, B. J., Zech, L., Cognition Technology Group at Vanderbilt. (1998). Smart environments that support monitoring, reflection, and revision. In D. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 305–346). Mahwah, NJ: Erlbaum.

Wang, S. (2005). *Online or paper: Does delivery affect results?* San Antonio, TX: Harcourt Assessment, Inc.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1):3–118.

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs*. San Antonio, TX: Harcourt Assessment.