SimScientists
WestEd

How Can Simulations Be Components of Balanced State Science Assessment Systems?

policy BRIEF

Edys S. Quellmalz, Matt D. Silberglitt, Michael J. Timms

Executive Summary

This policy brief reports on how simulation-based science assessments can become transformative components of multilevel, balanced state science assessment systems. Based on a six-state Enhanced Assessment Grant study of the technical quality, feasibility, and utility of SimScientists simulation-based science assessments, the brief recommends two possible models that policymakers may consider for incorporating these assessments into state science assessment systems.

The technical quality, feasibility, and instructional utility of the simulation-based assessments were evaluated in expert reviews by the American Association for the Advancement of Science (AAAS), in cognitive laboratories, and in the analyses of classroom pilot tests with 55 teachers and 5,465 students from 3 states, 28 districts, and 39 schools.

Data from this large pilot study provided evidence that the SimScientists benchmark assessments are of high technical quality, suitable for inclusion in a multilevel state accountability system. The pilot testing also documented the feasibility of implementation of the simulation-based assessments by large numbers of teachers across a wide variety of settings and technical infrastructures in districts and schools.

An external evaluator — the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) — found that students were highly engaged in the SimScientists assessments and able to complete them successfully. Teachers reported that the assessments were very useful for understanding student progress, measuring state standards, and adjusting instruction. Teachers and students believed the simulations had greater benefits than traditional paper-and-pencil tests because of the simulations' instant feedback, interaction, and visuals.

Analyses indicated that the simulation-based assessments met acceptable standards for reliability and validity. The study also found evidence that because the simulation-based assessments contain more visual representations and less text, they allow English language learners and students with disabilities to better demonstrate their science content knowledge and particularly their science inquiry skills than on more traditional paper-and-pencil tests.

A six-state Design Panel recommended two models states might consider for integrating such science simulations into their assessment systems. In the "Side-by-Side" model (see Figure 6 on page 9), states would aggregate the results of the simulation-based benchmark assessments collected over the school year and report them alongside state science reporting categories. This model could enhance *continuity* of the state science assessment system by adding multiple, continuous measures for the topics assessed in the science curriculum. These data would make the assessment system more *comprehensive* by increasing the coverage and detail of diagnostic information available on science achievement, and by permitting reports of proficiency in the three science content areas (life, physical, and Earth), topics within them (e.g., cells, ecosystems), and on inquiry practices (e.g., designing and conducting investigations). Unit benchmark reports would add to the *coherence* of classroom assessments and the state test reports.

The second, "Signature Task" model (See Figure 7 on page 10), involves use of the specifications and simulation environments to develop a pool of parallel tasks for states and/or districts to administer in a matrix sampling design. *Coherence* between classroom and state-level testing would be achieved by using the same specifications and simulation environments in the design of classroom assessments as in a statewide pool of signature tasks. *Continuity* and *comprehensiveness* would be improved, as in the Side-by-Side model, by adding the simulation-based data to the state science report.

Either of the two models would contribute to a balanced, multilevel state science assessment by providing: (1) a coherent set of nested, articulated assessments at the classroom, district, and state levels; (2) more comprehensive coverage of core standards; and (3) continuity of multiple measures collected during the year(s).

Findings from this Enhanced Assessment Grant study support the credibility of the SimScientists assessments for augmenting evidence of student achievement in classroom, district, and state levels of a state science assessment system. The SimScientists assessments can add rich, deep assessments of complex science learning and inquiry practices not typically measured adequately by traditional assessments.

Background

State science assessment systems are engulfed in a sea change. New science frameworks and standards call for deeper understanding of dynamic science systems and uses of science inquiry practices. Many states recognize that traditional assessment formats cannot adequately assess these aspects of science. States currently administer large-scale science assessments to all students at least three times between grades 3 and 12. While the majority of these assessments use traditional paper-based formats of multiple choice and constructed response, many states now recognize that important aspects of science are not assessed well in these formats. For example, knowledge of causal, temporal, and dynamic relationships among components within physical, life, and Earth systems, as well as inquiry processes, such as conducting investigations and communicating results, are difficult to test with traditional item formats. Some states have tested inquiry skills with hands-on performance assessments, but there are many logistical and economic challenges related to equipment, implementation, and scoring of such assessments both in classrooms and on the large scale required for state testing (Sausner, 2004).

To date, computer technologies have been used mainly to address the logistics of administration and scoring of assessment programs, but computers are beginning to show promise for the development of measures of complex learning useful for instruction and policy (Quellmalz & Pellegrino, 2009). 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. A new generation of assessments is showing potential to transform what, how, when, where, and why assessment occurs and how it can support teaching and learning.

Increasingly, computer technologies allow representations of domains, systems, models, data, and their manipulation in ways that previously were not possible. Dynamic models of ecosystems or molecular structures help scientists visualize and communicate complex interactions. This move from static to dynamic models has changed the nature of inquiry among professional scientists as well as the way that academic disciplines can be taught. Technology can also support the design of complex, interactive tasks that extend the range of knowledge, skills, and cognitive processes that can be assessed (Quellmalz & Haertel, 2004). For example, computer-based simulations can assess and promote understanding of complex systems by superimposing multiple representations and permitting manipulation of structures and patterns that otherwise might not be visible or even conceivable. Simulation-based assessments can probe basic foundational knowledge such as the functions of organisms in an ecosystem and, more importantly, they can probe students' knowledge of how components of a system interact and give students opportunities to investigate the impacts of multiple variables changing at the same time (Quellmalz, Timms, & Buckley, 2010).

The relatively small sample of knowledge and abilities captured by current traditional assessment provides an incomplete picture of student achievement in science. Many states are looking for new, innovative formats that can capture students' abilities to understand science systems and to use scientific inquiry. States are also looking for ways to build coherent, nested systems of assessment that augment state and district tests with end-of-unit, summative assessments and with curriculum-embedded formative assessments that improve learning. One way to satisfy the desire for assessments of learning about science systems and inquiry practices is to incorporate dynamic animations and interactive simulations of scientific phenomena, delivered over the Internet, through schools' networks and hardware. These simulation-based assessments have the potential to become credible components of state science assessment systems.

To explore the potential of simulation-based assessments, the U.S. Department of Education, Office of Elementary and Secondary Education funded an Enhanced Assessment Grant study on "Integrating Simulation-Based Science Assessments into Balanced State Science Assessment Systems." This collaboration included six states (led by Nevada, and including Connecticut, Massachusetts, North Carolina, Utah, and Vermont); WestEd; and CRESST at the University of California, Los Angeles. The collaborative studied the suitability of simulation-based science assessments developed by WestEd's SimScientists project as components of state science assessment systems. Science leaders from the six states formed a Design Panel to monitor the project and its implications for their state science assessment systems. Three states (Nevada, North Carolina, and Utah) pilot tested the SimScientists assessments that focused on two middle school science topics — ecosystems, and force and motion. For each topic, simulation-based, curriculum-embedded assessments provided opportunities for classroom-level formative assessment, offline reflection activities reinforced and extended the targeted

FIGURE 1. Student model for ecosystems, including model levels, content targets, and science practices

Model Level	Model Level Description	Content Targets by Model Level	Science Practices by Model Level
Component 	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
Interaction 	How do 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.
Emergent 	What is the overall behavior or property of the system that results from many interactions following specific rules?	Interactions 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 interaction among its components.

concepts and inquiry skills, and simulation-based unit benchmark assessments provided summative proficiency data.

Design Foundations of the SimScientists Assessments

The SimScientists simulation-based assessments were developed by WestEd in accordance with a strong set of assessment, pedagogical, and technological design principles that are described below.

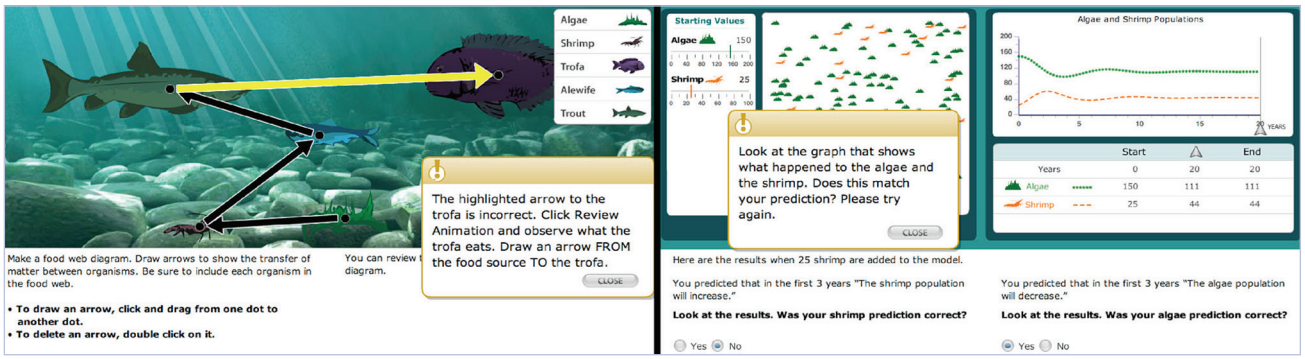
Evidence-centered design. The SimScientists assessments were developed using the evidence-centered assessment design approach which involves relating a model of student knowledge and skills to be assessed to a task model that specifies features of the task and questions to elicit evidence of learning, then to an evidence model specifying how proficiency is analyzed and reported (Messick, 1994; Mislevy & Haertel, 2007).

Model-based learning. The science concepts and practices assessed by the SimScientists assessments are based on national frameworks (including the draft Next Generation Framework for Science Education) and state standards which were used to develop assessment targets that reflect research on model-based learning (see www.simscientists.org). Rather than focusing on discrete factual content, the SimScientists assessments target connected knowledge structures that organize concepts and principles into features that are common to all systems — components, interactions, and emergent behaviors. Figure 1 presents the three generic levels of a system model applied to standards for middle school ecosystems and provides further detail on the particular content and science practices to be assessed.

Universal Design for Learning. The SimScientists assessments make use of the flexibility provided by digital technologies and recommended in the Universal Design for Learning framework (CAST, 2008). The visual, dynamic, and interactive features of simulations make assessment tasks more accessible to a greater range of students (Pellegrino, Chudowsky, & Glaser, 2001). SimScientists assessments offer audio and zoom accommodations to increase accessibility for some students who need accommodations. In addition, the assessments can be segmented to allow students to take more than one class period to complete them.

Assessment for learning. The effectiveness of formative assessment depends on several factors, including alignment of assessments with state standards, the quality of feedback provided to students, involvement of students in self-reflection and improvement, and whether teachers actually make adjustments to their instruction

FIGURE 2. SimScientists embedded assessments provide feedback and coaching



based on the assessments (Black & Wiliam, 1998). The SimScientists assessments include features that address all of these factors.

Description of the SimScientists Assessments

For each topic, SimScientists provides *embedded assessments*: two for ecosystems, three for force and motion. The embedded assessments are designed to be given at appropriate points when a sub-topic has been covered in the regular classroom instruction. Teachers are given tools to align the assessments to the learning goals in their classrooms, allowing them to embed the assessments at appropriate times in their curricula. During the embedded assessments, students complete tasks such as making observations, running trials in an experiment, recording data, interpreting data, making predictions, and explaining results. They answer questions by various methods such as selecting from a choice of responses, changing the values of variables in the simulation, drawing arrows to represent forces, and typing explanations. Throughout the tasks, the system gives students feedback and graduated levels of coaching so that students have multiple opportunities to confront their misconceptions, with increasing scaffolding based on the amount of help needed. Figure 2 presents screenshots of two SimScientists embedded assessments that provide immediate feedback and coaching as students interact with the simulations. These screenshots show the kinds of signature tasks that can serve as templates for components of classroom, district, and state assessments.

The left screenshot shows a task that asks students 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 the student to observe again by reviewing the animation and to draw the arrow from food source to consumer. Feedback also addresses common misconceptions. The right screen shot shows feedback and coaching for an investigation of population changes.

The system also provides a progress report to students at the end of each embedded assessment. Because reports in the form of grades can undermine learning and student motivation, each SimScientists embedded assessment provides a report with informative feedback that helps students connect their success in the assessment to their effort (Covington, 1999; Maehr & Midgley, 1996).

Each embedded assessment also provides a report to the teacher that includes classifying each student into one of three groups based on the amount of feedback and coaching students received on different parts of the content. Each embedded assessment is followed by offline classroom *reflection activities* in which teachers organize students into the recommended groups and students complete tasks that provide opportunities for reflection and improvement. These reflection activities along with the assessment's detailed reports about students' progress on each assessment target provide teachers with the tools to adjust instruction to match students' needs based on the results of the assessments. Figures 3 and 4 show examples of the kinds of reports that students and teachers can use in the process of formative assessment.

In addition to embedded assessments, there are SimScientists *benchmark assessments*, which students take at the end of curriculum units. Each benchmark assessment is a compilation of the embedded assessment activities, transferred into a new context. For example, the embedded assessments for the ecosystems topic present a lake ecosystem (see Figure 2); the benchmark assessment uses the same activities, but the setting is a grasslands ecosystem

FIGURE 3. Embedded assessment individual report to student

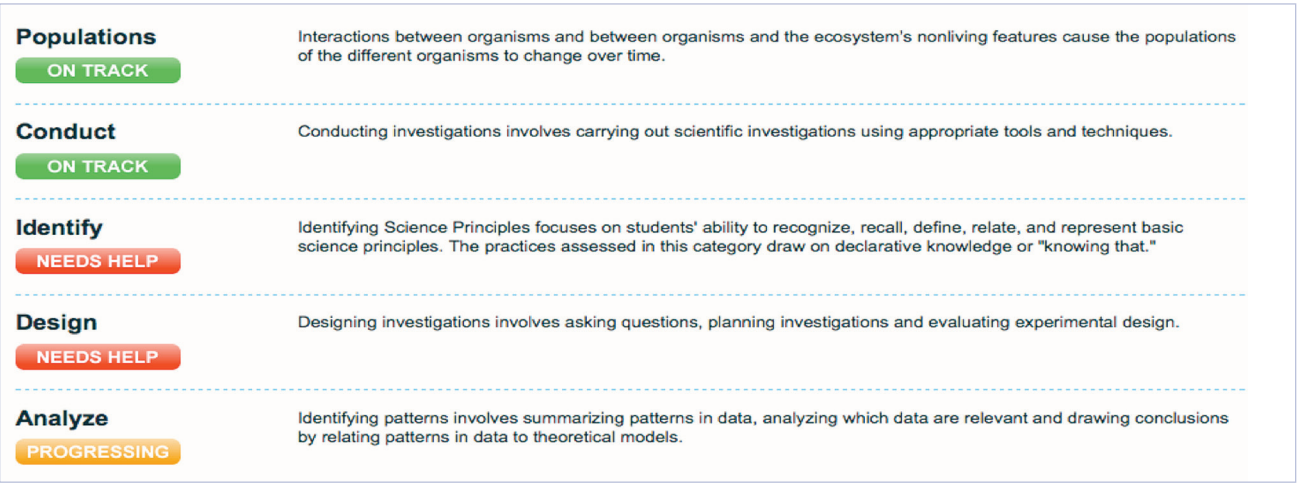


FIGURE 4. Embedded assessment class progress report to teacher

Content		NH Needs Help	P Making Progress	OT On Track
▸ Populations	NH PROG OT	3 (11%)	10 (36%)	15 (54%)
Inquiry				
▸ Conduct	PROG OT	1 (4%)	11 (39%)	16 (57%)
▸ Identify	OT	1 (4%)	3 (11%)	24 (86%)
▸ Design	PROG OT	2 (7%)	13 (46%)	13 (46%)
▸ Analyze	PROG OT	0 (0%)	14 (50%)	14 (50%)

with different organisms and different, though parallel relationships between these organisms. In this way, students cannot simply memorize the material from the embedded assessments, and instead have to show that they can transfer their knowledge and inquiry skills. The other major difference between the embedded assessments and the end-of-unit benchmark assessment is in the use of feedback and coaching. The embedded assessments provide feedback and coaching to scaffold students' learning, but the benchmark assessment does not provide any coaching. Upon students' completion of the benchmark assessment, the teacher uses the SimScientists assessment system to score the written responses, then those scores, along with the scores from machine-scored tasks, are evaluated by the system to produce summative reports for the students and the teacher on achievement on the state science standards and on specific content and inquiry targets addressed in the unit (see Figure 5).

The Enhanced Assessment Grant Study

The goals of the Enhanced Assessment Grant study of SimScientists assessments were to establish the assessments' technical quality, feasibility in the classroom, and effects on student performance (especially for English language learners and students with disabilities), and to propose alternative models for integrating simulation-based assessments into state assessment systems. The SimScientists assessments were tested in three phases that built upon one another: first in cognitive laboratories in which students were asked to think aloud as they worked through the assessments, then in feasibility tests in classrooms to ensure that the assessments worked in school settings, and lastly in a large-scale pilot test to collect data on the technical quality of the assessments. Cognitive laboratory sessions were conducted with 28 individual middle school students and 4 teachers during development to provide preliminary evidence of usability and construct

FIGURE 5. End-of-unit benchmark reports to teacher for class and individual students

validity. Results from the cognitive laboratory sessions also informed revisions made to the assessments during their development. Feasibility tests in two classrooms, one for each topic, provided information on the logistical challenge of delivering assessments via the Internet and led to changes to how the assessments were transmitted to the classroom, as well as showing that the length of the assessments was correct for class periods in the schools.

The pilot study, which took place in spring 2010, included 55 teachers and 5,465 students in three states, 28 districts, and 39 schools. 3,529 students took part in the test of the ecosystems assessments, and 1,936 students tested force and motion. During the administration, student response data were collected from the SimScientists assessments and also from a posttest composed of conventional multiple-choice items on the same topics drawn largely from an American Association for the Advancement of Science (AAAS) bank of calibrated items and supplemented with items developed by WestEd. In addition, student data were collected, including gender, ethnicity, English language learner status, and whether students had an Individualized Education Plan (IEP) or Section 504 Accommodation plan.

Teacher data were collected through surveys, interviews, and classroom observations. Teacher surveys asked about their curricula, the feasibility of the assessment system, the utility of the reports, and students' opportunity-to-learn. The pilot study collected computer logs that recorded students' use of the assessments and teachers' use of the learning management system (LMS) that is used to deliver the assessments.

Main Findings

The SimScientists assessments achieved acceptable standards of reliability and validity.

A psychometric analysis of the student responses on the ecosystems benchmark assessment showed that all except one of the 45 items were contributing information relevant to the overall measure of science content and practices. The reliability was .76 for the ecosystems benchmark assessment, which is considered acceptable (George & Mallery, 2003), particularly for an assessment that uses simulations and has a mixture of item types, including selected response items, and short written responses scored by teachers. Similarly, for the force and motion benchmark assessment, all except one of the 41 items fitted the measurement model, which indicated that all items except one were contributing information relevant to the overall measure. The reliability for the force and motion benchmark assessment was .73, which is acceptable.

Evidence of validity of the simulation-based assessments came from several sources. A review by content experts at AAAS confirmed that the assessment tasks were aligned to important content and inquiry targets as defined by the standards. An analysis of think-aloud sessions with 28 students showed that they were applying

TABLE 1. Comparison of gaps between the performance of English language learners and of students with disabilities versus performance of the general population on the simulation-based benchmark assessments and traditional posttests

Group	Ecosystems Posttest	Force & Motion Posttest	Ecosystems Benchmark	Force & Motion Benchmark
ELL Students	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)

the intended content and inquiry skills an average of 84 percent of the time as they worked through the tasks, indicating that the assessments did elicit the targeted knowledge. Further validity evidence came from a correlation of the student performance on the science content and inquiry measures of 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 practices but that the measures were not exactly the same. This was expected because 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 skills were lower than the correlations for content knowledge, supporting this interpretation.

The study also found that the benchmark assessments distinguished student performance on inquiry practices more effectively than the posttests. The correlation of the content and inquiry dimensions on the posttest for ecosystems (.85) and force and motion (.92) were higher than those for the benchmark assessments (.70 and .80, respectively). This indicates that the distinction between the measures of content and inquiry is greater in the simulation-based benchmark assessments than on the traditional items of the posttests.

English language learners and students with disabilities performed better on the simulation-based assessments.

Overall, students performed better on the benchmark assessments than on the more conventional posttests, and performance gaps between both English language learners (ELL students) and students with disabilities compared with other students were reduced on the benchmarks. To determine the effect of the simulation-based assessments on ELL students and students with disabilities, their performances on the benchmark assessments were compared with their performances on the conventional posttests. Table 1 compares performance gaps of ELL students and students with disabilities compared with a reference group of students who are neither ELL students nor students with disabilities. Although the average performances of ELL students and students with disabilities on the SimScientists benchmarks are lower than those of the reference group, the gaps between the focal groups and the reference group are comparatively smaller than for the posttests. For ELL students, the performance gap on the benchmarks averaged 12.1 percent compared with 25.7 percent on the posttests. Similarly, the gap for students with disabilities on the benchmarks averaged 7.7 percent compared with 18.0 percent on the posttests. This evidence suggests that the multiple representations in the simulations and active manipulations may have provided alternative means, other than written text, for ELL students and students with disabilities to understand the assessment tasks 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. For ELL students, the performance gap on the benchmarks averaged 8.8 percent compared with 30.4 percent on the posttests. Similarly, the gap for students with disabilities on the benchmarks averaged 5.9 percent compared with 22.9 percent on the posttests. Again, this evidence indicates that ELL students and students with disabilities were

TABLE 2. Comparison of gaps in inquiry skills between performance of English language learners and of students with disabilities versus performance of the general population on the simulation-based benchmark assessments and traditional posttests

Group	Ecosystems Posttest	Force & Motion Posttest	Ecosystems Benchmark	Force & Motion Benchmark
ELL Students	25.6% (n=123)	35.1% (n=50)	6.6% (n=126)	10.9% (n=50)
Students with Disabilities	25.5% (n=183)	20.3% (n=153)	5.6% (n=189)	6.2% (n=153)

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

The simulation-based assessments were feasible in the classroom, engaging to students, and provided useful information about learners.

Joan Herman at CRESST at the University of California, Los Angeles, conducted an independent evaluation of SimScientists assessments. Teacher surveys collected by WestEd and CRESST and classroom observations by CRESST indicated that nearly all the teachers were able to successfully administer the assessments online using the existing infrastructure in each school. Students were able to complete the simulation-based assessments during the allotted class periods and to use the interfaces to complete the assessment tasks. The evaluation case study report stated "students were active and engaged during the assessments and able to use the computer assessments effectively." In computer labs, teachers introduced the assessments, then monitored student progress, providing assistance as needed. During the reflection activities, teachers introduced the activities, monitored the groups, oversaw the merger of small groups with larger ones, and oversaw the presentations.

The CRESST evaluation report summarized data from case study observations, teacher surveys, and interviews as follows:

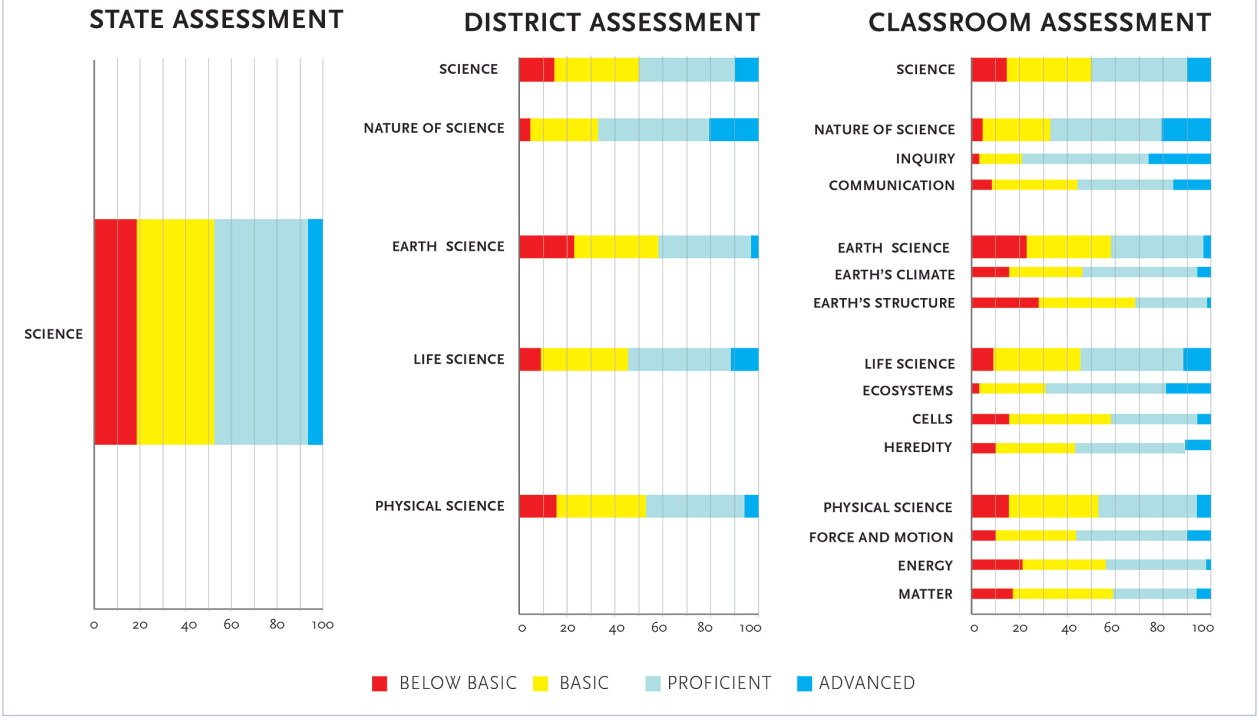
> *Teachers reported that the embedded assessments were very useful for understanding student progress and adjusting their instruction. Teachers and students believed that the simulations had greater benefits than traditional paper-and-pencil tests because of the simulations' instant feedback, interaction, and visuals. The instant reports allowed teachers to easily see which questions students had the most difficulty with so that they could tailor their lessons accordingly. Teachers agreed that the assessments would be useful in measuring their individual state standards. (Herman et al., 2010)*

There are two potential models for integrating simulation-based assessments into a balanced state science assessment system.

A balanced state assessment seeks to have assessments at the classroom, district, and state levels that are mutually reinforcing. The combination of assessments should be **coherent**, **comprehensive** in coverage of state science standards, and provide **continuity** of assessments through multiple forms and occasions. The Enhanced Assessment Grant study formulated two alternative models states could use to incorporate simulation-based science assessments: a *Side-by-Side* model that involves using the unit benchmark assessment proficiency data to augment state reports; and a *Signature Task* model that involves using simulation-based tasks in parallel to those in the benchmarks as part of state or district tests.

Figure 6 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 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

FIGURE 6. Side-by-Side model, showing how data reported from unit benchmark assessments can augment information from district and state science reports



"through- course" measures that are directly linked in time and substance to units on science systems such as climate or Earth's structure.

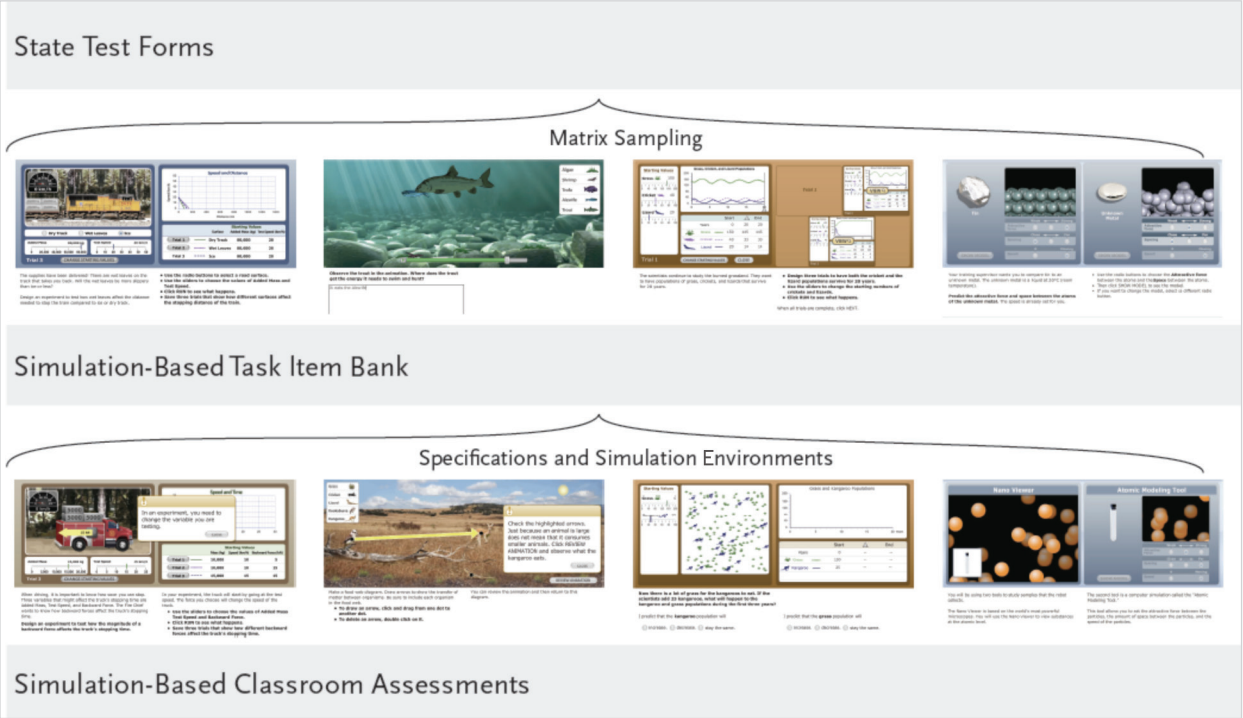
Figure 7 portrays the Signature Task model in which states and districts draw upon the specifications and rich simulation environments developed for the classroom-level unit benchmark assessments to create a new, parallel set of key, or signature, tasks such as drawing a food web or conducting a predator-prey 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 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. In Figure 7, 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 nested assessments collected from different levels of the education system.

Significance

The simulation-based assessments studied in this Enhanced Assessment Grant could contribute to the coherence, comprehensiveness, and continuity of states' science assessment systems. **Comprehensiveness** would be improved by using simulation-based unit assessments to add measurements of science standards for integrated system knowledge and active inquiry practices. **Continuity** would be improved by the multiple measures that unit benchmark assessments could add to state science assessment reports. **Coherence** could be forged by a nested set of simulation-based assessments in the form of curriculum-embedded modules for formative

FIGURE 7. Signature Task model, showing how parallel tasks can be developed for state and classroom assessments



uses, unit benchmark assessments for summative proficiency, and use of the unit benchmark data or tasks in district or state science testing.

Technical quality. The high degree of reliability on the simulation-based assessments provided evidence of the technical quality of the assessments. These technical quality data are particularly important, given the wide range of item formats — from more traditional multiple-choice and constructed-response items to innovative and interactive items, including machine scoring and teacher scoring. Further evidence of technical quality is provided by the results of think-alouds which demonstrate that the items elicited the intended content knowledge and inquiry abilities. In addition, validity was documented by expert reviews of the alignment of the assessments to national and state standards in science.

Feasibility. The successful implementation of the SimScientists assessments across a diverse range of schools and districts demonstrates the feasibility of such assessments. Our sample included large urban settings, small rural schools, charter schools, and a juvenile detention facility. We demonstrated the feasibility of state assessment systems with innovative formats and rich, dynamic stimuli that can assess a broader range of knowledge and skills in science.

Utility. Evidence of utility from observations, surveys, and interviews indicates that the SimScientists assessment system composed of embedded, formative assessments and summative unit benchmarks helps students understand their own strengths and weaknesses in science. Teachers found that the embedded assessments provide useful information for monitoring student progress and for adjusting subsequent instruction. The overwhelmingly positive responses to the formative components of the system for improving student learning and to the summative components for providing information to teachers demonstrate the utility of the SimScientists assessment system.

Conclusion and Implications

The Enhanced Assessment Grant is one of the first studies to provide research-based evidence that systematically developed and verified simulation-based science assessments used for formative and summative purposes can achieve high technical quality, be broadly implemented, and have strong instructional utility. The evidence provides support for the claims that innovative, technology-enhanced assessments can be credible components of multilevel, balanced state science assessment systems.

References

AAAS. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

Black, P. & Wiliam, D. (1998) Assessment and classroom learning. *Assessment in Education*, *5*, 7–74.

Buckley, B. C. (2000). Interactive multimedia and model-based learning in biology. *International Journal of Science Education*, *22*(9), 895–935.

CAST. (2008). *Universal design for learning guidelines, version 1.0.* Wakefield, MA: Author.

Covington, M. V. (1999). Caring about learning: The nature and nurturing of subject-matter appreciation. *Educational Psychologist*, *34*(2), 127–136.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Herman, J., Dai, Y., Htut, A. M., Martinez, M., & Rivera, N. (2010). *CRESST evaluation report: Evaluation of the Enhanced Assessment Grant (EAG)*. Los Angeles: CRESST.

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

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

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

NAGB. (2008). *Science framework for the 2009 National Assessment of Educational Progress.* Washington, DC: Author.

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

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

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

Quellmalz, E. S., & 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. S., Timms, M., & Buckley, B. (2010). Science simulations for assessment. *International Journal of Learning Technology*, *5*(3).

Sausner, R. (2004, August). Ready or not: NCLB English and math requirements. *District Administration.* Retrieved 5/29/2007 from http://districtadministration.ccsct.com//page.cfm?p=832.

U.S. Department of Education, National Center for Education Statistics. (2011). *NAEP Data Explorer.* Available at http://nces.ed.gov/nationsreportcard/naepdata/.

This policy brief is based on work supported by the U.S. 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 brief are those of the authors and do not necessarily reflect the views of the U.S. Department of Education or the National Science Foundation. Additional publications from this study can be found at http://simscientists.org.

For more information on the issues covered in this Policy Brief, contact Edys Quellmalz, Director of Technology Enhanced Assessment and Learning Systems, 400 Seaport Ct, Suite 222, Redwood City, CA 94063-2767; phone 650.381.6427; email equellm@WestEd.org.

WestEd, a nonprofit research, development, and service agency, works with education and other communities to promote excellence, achieve equity, and improve learning for children, youth, and adults. While WestEd serves the states of Arizona, California, Nevada, and Utah as one of the nation's Regional Educational Laboratories, our agency's work extends throughout the United States and abroad. It has 16 offices nationwide, from Washington and Boston to Arizona, Southern California, and its headquarters in San Francisco.

For more information about WestEd, visit our website: WestEd.org; call 415.565.3000 or, toll-free, (1.877) 4-WestEd; or write: WestEd / 730 Harrison Street / San Francisco, CA 94107-1242.

© 2011 WestEd. All rights reserved.

730 Harrison Street
San Francisco, CA 94107-1242

Address service requested