Paper #1
Abstract Title Page

Title: Multilevel Assessments of Science Standards
Author(s): Edys S. Quellmalz[1], Michael J. Timms[2], Matt D. Silberglitt[2]
Affiliations of authors: 1: WestEd, 400 Seaport Court, Redwood City, CA 94063; 2: WestEd, 300 Lakeside Dr., 25th Floor, Oakland, CA 94612
Author's emails: equellm@wested.org, mtimms@wested.org, msilber@wested.org
Contact email for paper: equellm@wested.org

Abstract Body

Background / Context:

The Multilevel Assessment of Science Standards (MASS) project is creating a new generation of technology-enhanced formative assessments that bring the best formative assessment practices into classrooms to transform what, how, when, and where science learning is assessed. The project is investigating the feasibility, utility, technical quality, and effectiveness of formative assessments, summative assessments and the Learning Management System (LMS) developed in the SimScientists program. The SimScientists simulation-based assessments present dynamic, engaging interactive tasks of established technical quality that test complex science knowledge and inquiry skills going well beyond the capabilities of print tests. The SimScientists curriculum-embedded assessments serve formative purposes in two ways: (1) to provide immediate feedback related to an individual student's performance, and (2) to offer links to additional instruction and coaching. The formative assessment process incorporates: *frequent* use of standards-based classroom assessments; feedback that is *timely*, *individualized*, and *diagnostic*; *supplementary instruction* that is individualized; and *self-assessment* and reflection activities that help students confront misunderstandings, make new connections, and become more reflective, self-regulating learners. Two to three embedded assessments have been created for a unit. The summative assessments were designed as end-of-unit, benchmark tests.

Purpose / Objective / Research Question / Focus of Study:

The MASS project is funded by IES and has the following goals:

- ☐ Use systematic design principles to create formative assessments with technical quality to be used during (embedded) and at the end of (benchmark) science curriculum units.
- ☐ Use systematic assessment principles to create a coherent, multilevel state science assessment system by aligning (1) the items within the embedded assessments and the items within the benchmark assessments with the student, task, and evidence models used to design them (horizontal alignment), and (2) the designs of the embedded and benchmark assessments and items with state science standards and relevant items on the state science test (vertical articulation).
- ☐ Study the relationship of the formative assessments and activities to student learning.
- ☐ Study the validity of the use of data from the embedded and benchmark assessments for interpreting student performance on the targeted science standards.
- ☐ Describe the components of the formative assessments and their implementation so that they can serve as scalable models.

Setting:

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, in cognitive laboratories, and in the analyses of classroom pilot tests with 55 middle-school teachers, from three states, 28 districts, and 39 schools.

Research on the instructional effects of the embedded assessments for formative use on the simulation-based summative unit benchmark assessments and a conventional post test is currently being conducted in the classrooms of five teachers and approximately 800 students. An additional five to eight teachers will participate in the fall of 2011.

Population / Participants / Subjects:

The pilot test involved 5,867 middle school students in Spring 2010. This population is from a range of small to large schools and districts, including rural, urban, and suburban districts,

a variety of ethnic and socioeconomic backgrounds, and includes English learners and students with disabilities. In the field test, approximately 800 middle school students are participating in Spring 2011; an additional 500-800 middle-school students will participate in Fall 2011. These populations are from a large school district, a variety of ethnic and socioeconomic backgrounds, and include English learners and students with disabilities..

Intervention / Program / Practice:

Three middle school topics are included in the project: Ecosystems, Force & Motion, and Atoms & Molecules. For each topic, a set of simulation-based curriculum-embedded assessments that provided immediate, individualized feedback and graduated coaching administered during an instructional unit offered opportunities for formative assessment. (please insert Figure 1 here) Off-line reflection activities reinforced the targeted concepts and inquiry skills and their transfer to novel contexts and also supported collaboration and scientific discourse. A simulation-based unit benchmark assessment at the end of each unit provided summative data on proficiency. Pre and post tests were administered comprised of traditional multiple choice items.

Research Design:

The development and pilot phases of the study used a mixed-methods design that included cognitive laboratories, teacher surveys and interviews, classroom observations, and logs of students' use of the assessments and teachers' use of the LMS. To study the effectiveness of the curriculum-embedded assessments and follow-up reflection activities, we employ an alternate treatments design, randomized within teacher. Table 1 shows the design of this phase of the study. (please insert Table 1 here)

Data Collection and Analysis:

During the development phases and pilot study, the SimScientists assessments were tested in three phases which built upon one another; cognitive laboratories in which students were asked to think-aloud as they worked through the assessments, feasibility tests in the classroom to ensure that the assessments worked in school settings, and 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 four teachers during development to provide preliminary evidence of usability and construct validity. Results from the cognitive laboratory sessions also informed revisions made to the assessments during their development.

During the pilot study, which took place in spring 2010, students took part in the test of the ecosystems and force & motion assessments. 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 AAAS bank of calibrated items and supplemented with items developed by WestEd. In addition, student demographic data were collected, including gender, ethnicity, if students were English Language Learners, and if they had an Individualized Education Programs (IEP) or Section 504 Accommodation plans.

Data were collected from teachers in the study 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 targeted content and inquiry skills. The pilot study collected computer logs recording students' performance on the assessments and teachers' use of the LMS used to deliver the assessments.

In the current phase of the project, in Spring and Fall 2011, students are taking part in a field test of the ecosystems and atoms & molecules assessments. During the administration, student response data are collected from the SimScientists assessments and also from a pretest

and a posttest composed of conventional multiple-choice items on the same topics. In addition student demographic data are being collected, including gender, ethnicity, if students were English Language Learners, and if they had an Individualized Education Programs (IEP) or Section 504 Accommodation plans.

Findings / Results:

In the pilot test, nearly all the teachers were able to successfully administer the assessments online using the existing infrastructure in the 39 schools and 28 districts in three states. In a small number of cases, however, "unique" network infrastructures presented greater challenges to implementation. A help line was available to support teachers in these circumstances, and in all but one case, teachers were able to implement the assessments after some additional troubleshooting.

On our reliability measures, items related to a common task, e.g., draw a food web, were bundled together and all bundles of items fitted the measurement model (infit between 0.8 and 1.2), which indicates that all the items were contributing information relevant to the overall measure. The reliabilities were 0.85 for the Ecosystems benchmark assessment and 0.79 for the Force & Motion benchmark assessment, which are very good for assessments that are a mixture of selected response, measures of use of the simulations, and short written responses scored by the teachers.

Science content measures from the benchmark assessment were significantly correlated with the posttest. All of the correlations were statistically significant, although they were moderate (.57 to .64). We expected only moderate correlations because the purpose of the simulation-based assessments was 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. Table 2 shows the correlations between posttest and benchmark ability estimates for content and inquiry. The analysis of the 28 think-aloud studies provided further evidence that the assessment tasks and items were eliciting evidence of the intended constructs. (please insert Table 2 here)

Overall, students performed better on the unit benchmark assessments than on the conventional posttest. This is indicated by the fact that the mean percent score on the benchmark versus the posttest was 70% vs. 55% on ecosystems and 67% vs. 53% on force & motion and, this difference in favor of the benchmark persists even when the items were analyzed using a partial credit item response model to scale the items on the same difficulty measure. As shown in Table 3, students who took both the benchmark and the posttest found the items easier on the benchmark, a difference of .54 logits on the ecosystems and .87 on the force and motion assessments. These differences represent .49 of the standard deviation on ecosystems and .82 of a standard deviation on force and motion. (please insert Table 3 here)

To determine the effect of the simulation-based assessments on English Language Learners (ELL) and students with disabilities (SWD) their performances on the benchmark assessments were compared to performance on the posttest of conventional items. Table 4 compares performance gaps of ELL and SWDs to a reference group of students who are neither English Language Learners nor students with disabilities. The table includes comparisons of performance gaps on the SimScientists benchmark assessments, the 30-item posttests used in the study, and the National Assessment of Educational Progress (NAEP) 2009 Science. 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 and for the 2009 NAEP Science. This evidence lends support to the

suggestion 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. (please insert Table 4 here)

The project's external evaluator, CRESST, conducted classroom case studies to examine implementation of the simulation-based assessments and found that students were highly engaged in the SimScientists assessments and able to complete them successfully. Teachers reported that the embedded assessments were very useful for understanding student progress and for 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. Teachers agreed that the assessments would be useful in measuring their individual state standards.

Conclusions:

The simulation-based assessments studied in this project could contribute to the coherence, comprehensiveness, and continuity of a state science assessment system. **Comprehensiveness** would be improved by using simulation-based unit assessments to add measurements of science standards for integrated system knowledge and active inquiry practices. **Continuit**y would be improved by the multiple measures 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 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 to innovative, interactive items, including machine scoring and teacher scoring, Further support is provided by the results of think-alouds, that demonstrate construct validity given that the items elicited the intended content knowledge and inquiry abilities. In addition, validity was also 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 the embedded assessment reports useful information sources for monitoring student progress and for adjusting subsequent instruction accordingly. The positive responses to the formative components of the system for improving student learning and the summative components for providing information to teachers demonstrates the utility of the system.

Benefits for Learning. The effects of the embedded assessments on unit benchmark assessments and the post test will be examined in the alternative treatments design being conducted in the field test in Spring 2011 and Fall 2011. The results of the Spring study for the ecosystem assessments will be reported.

Appendices

Appendix A. References

Appendix B. Tables and Figures

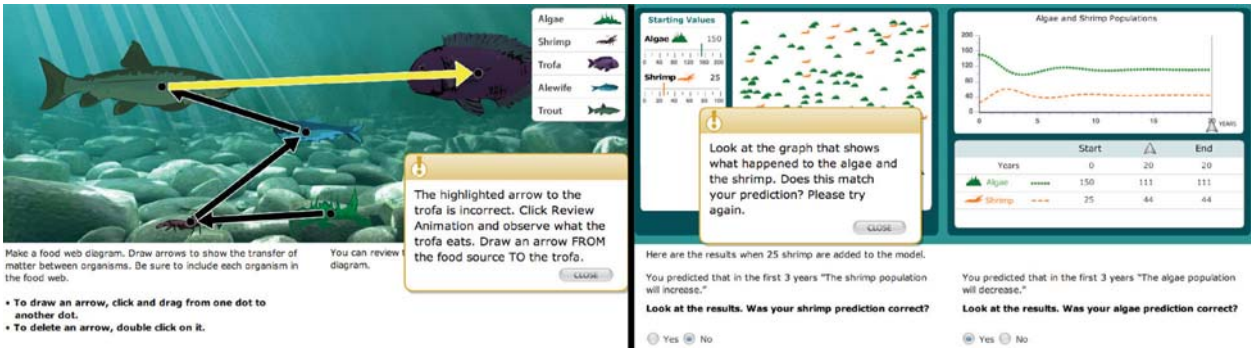

Figure 1. SimScientists embedded assessments provide feedback and coaching

Teacher #		Teacher's regular instruction							
Section # Treatment A	pretest		Embedded Assessment & Reflection Activity		Embedded Assessment & Reflection Activity	posttest	Benchmark		
Section # Treatment B	pretest		Embedded Assessment & Reflection Activity		Embedded Assessment & Reflection Activity	Benchmark	posttest		
Section # Treatment C	pretest					posttest	Benchmark	Embedded Assessment & Reflection Activity	Embedded Assessment & Reflection Activity
Section # Treatment D	pretest					Benchmark	posttest	Embedded Assessment & Reflection Activity	Embedded Assessment & Reflection Activity

Table 1: Alternate treatment design

	Ecosystems (n=2924)		Force & Motion (n=1496)	
	Content Ability estimate	Inquiry Ability estimate	Content Ability estimate	Inquiry Ability estimate
Correlation with ability estimates on posttest	.64**	.57**	.61**	.60**

*Table 2. Correlations of Benchmark to Posttest **Correlation is significant at the 0.01 level (2-tailed).*

	Benchmark	Posttest	Difference
Ecosystems Mean difficulty (logits)	-0.83	-0.29	.54
Force & Motion Mean difficulty (logits)	-0.87	0	.87

Table 3. Comparison of the average difficulty of the benchmarks and posttests

Total	NAEP Average	Ecosystems Posttest	Force & Motion Posttest	Ecosystems Benchmark	Force & Motion Benchmark
English Language Learners	16.8%	24.0%	27.4%	10.6%	13.6%
Students with Disabilities	11.7%	20.2%	15.7%	8.4%	7.0%

Table 4. Comparison of gaps in total performance between performance of English Language Learners and students with disabilities and the general population on 2009 NAEP Science and on the simulation-based benchmark assessments and static posttests.