Developing a GLOBE Assessment in North Carolina

William R. Penuel & Valerie M. Crawford
Center for Technology in Learning
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493

SRI International (SRI) has provided evaluation services for the GLOBE program since 1995. A key component of each year's evaluation has been the assessment of students' science knowledge, ability to work with data, conceptions of the practice of science, and environmental awareness. In 2000-01, SRI worked closely with staff at the North Carolina Department of Public Instruction on pilot tests of student learning in 8th grade science.

GLOBE and North Carolina's Science Standards

SRI's task was to develop assessment items at the intersection of GLOBE objectives and North Carolina's own objectives for student learning in middle school science. North Carolina revised its standards for K-12 science education in 2000 and formally approved them in 2001. The new standards place a strong emphasis on developing students' understanding of science as a process of inquiry advanced by new technologies and their understanding of how science can be useful in making personal and social decisions. Content standards for each grade level are focused around a single "big idea" in science, and students are expected to master just 4-5 content areas at each grade level.

North Carolina wrote its middle school standards to correspond closely with goals for three GLOBE investigation areas: Soils (6th grade), Atmosphere (7th grade), and Hydrology (8th grade). In addition, eighth graders in North Carolina are expected to learn how to interpret data from satellite images of Earth, which is part of the Land Cover/Biology Investigation Area. Since the new standards were first proposed, trainers from GLOBE's 10 U.S. partners in North Carolina and Clara Stallings, science consultant with the North Carolina Department of Public Instruction, have been active in preparing teachers to implement GLOBE and encouraging teachers to use GLOBE to help meet the middle school science standards. Teachers interviewed by SRI in North Carolina see GLOBE as an important resource not only for teaching the content standards but also for helping students develop a deeper understanding of scientific inquiry.

There are some areas where North Carolina's science standards and GLOBE do not overlap. Some content areas are not covered by GLOBE, such as motion and forces and genetics and heredity. Even in content areas covered by GLOBE, there are some concepts that students would not necessarily learn by following GLOBE protocols or implementing learning activities from the GLOBE Teacher's Guide. For example, in eighth grade, students are expected to study ocean ecosystems as part of their study of the hydrosphere. Students are expected to examine relationships among different organisms in the ocean, a topic that GLOBE does not address directly.

Our Assessment Focus: Hydrology and Land Cover

Researchers at SRI met with Clara Stallings of the North Carolina Department of Public Instruction (DPI) in early fall 2000 to identify key goals for our assessment development process. We decided to develop measures of student learning of GLOBE Hydrology and Land Cover concepts and skills that were part of the eighth grade science curriculum in North Carolina. Our plan for 2000-01 was to develop and pilot test items and to use the assessment tool developed to evaluate program effectiveness in 2001-02. The North Carolina DPI hoped to be able to use some items for end-of-year content tests in the middle grades and other items for formative use in the classroom by teachers. SRI hoped to learn more about the effectiveness of GLOBE when there is strong statewide support for the program and when state standards are closely aligned with GLOBE's own objectives.

Through our assessment piloting process, we aimed to identify a set of items that would be sensitive to the effects of GLOBE participation in the context of a GLOBE-aligned science curriculum; that is, items where active study of GLOBE concepts and participation in inquiry focused about environmental science data could detect a significant advantage for GLOBE participants. In addition to testing students in North Carolina, we tested a stratified random national sample of active GLOBE classrooms and a random sample of classrooms where teachers had been trained in GLOBE but were not yet implementing GLOBE at the same time we administered post-tests to students in North Carolina. Testing these groups of students permitted us to compare the effects of GLOBE in a GLOBE-aligned science curriculum (North Carolina sample), with the effects of GLOBE in non-aligned curricula (the national sample), so we could better understand how the variable of "curriculum alignment" affects student performance on our measure.

The Assessment Development Process

To develop the assessment instrument, we were guided by an evidence-centered (Messick, 1992; Mislevy et al., 2000) approach to assessment design. Identifying what kind of evidence is needed to infer students have mastered a clear set of learning objectives is the heart of this approach to development. Our approach also involved significant input from DPI staff in North Carolina, GLOBE scientists, and other SRI assessment experts, to augment the validity of our assessment instrument.

Gathering Existing Assessment Items. In Year 2, SRI had created a set of assessment items that cut across GLOBE investigation areas as part of our evaluation efforts. We included in the current pool of possible items those items focusing on Hydrology and Land Cover/Biology with significant *p* values, that is, items where GLOBE students scored significantly higher than non-GLOBE students. We also consulted the Center for Technology in Learning's own assessment group, the Internet, and the TIMSS and NAEP databases for possible items.

Developing a Content Coverage Matrix. After an initial set of items had been identified, we developed a matrix showing objectives for student learning that are common between GLOBE objectives and North Carolina's science standards. The objectives listed included both Hydrology and Land Cover/Biology concepts and two aspects of the inquiry process, planning an investigation and analyzing data.

Creating New Items and Scoring Rubrics. We determined alignment of each item with the content coverage matrix. This also helped us identify where we

needed additional items to measure each objective. Our goal was to develop at least 3 items for each major objective, using three different formats: multiple choice, short-answer, and long-answer open-ended items. We developed new items and initial scoring rubrics to meet our goals for coverage. Because we knew we would end up with fewer items for each objective than we piloted, we developed 5-6 items for each objective.

Expert Content Review. We asked GLOBE scientists within the Hydrology and Land Cover/Biology Investigation Areas and staff at the GLOBE office to review items for accuracy of content. We revised our items again based on their comments and feedback.

Pre-Piloting Items. In late fall 2000 we pre-piloted a set of items with six classes. For each classroom, we timed how long it took students to complete the test; in some classes, students were timed on each item (working in pairs, alternating time-keeping and item-completing). We alsoconducted think-aloud protocols with students at high, medium, and low reading abilities to determine how they were approaching solving particular problems on the assessment, to identify any linguistic or other barriers, instructions that were unclear, and items that students interpreted in ways divergent from the item purpose.

Revising Items. Items were revised again based on the classroom piloting and results of the think aloud protocols.

Online Form Piloting. An online test form was created. Researchers and programmers tested the online form performance and usability.

Pilot Testing Items. In late February and early March, we piloted the items in 5 classrooms in North Carolina using the online assessment form. The teachers in these classrooms were all GLOBE-trained teachers. None had implemented GLOBE, but they all planned to implement GLOBE in the next 3 months with their classes. In May, we asked these same classrooms to complete post-tests, though only 2 classrooms completed the assessment again. The assessments were administered in 5 randomly-selected active GLOBE classrooms and 3 randomly-selected classrooms where teachers had been trained in GLOBE but had not implemented GLOBE.

Analyzing Pilot Results. We analyzed data from student assessments and student surveys to examine which items were instructionally sensitive for two scales: hydrology content and science inquiry skills. We were unable to identify a comparable “active” Land Cover/Biology sample, so we decided to focus our initial analysis on results from a national sample of active reporters of the Hydrology protocols.

Pilot Test Results

To test which items successfully discriminated between active GLOBE classrooms and non-GLOBE classrooms, we conducted an independent-samples *t*-test for each item in our two scales. For the hydrology content scale, GLOBE students scored significantly higher on 5 of 13 items ($.0005 < p < .05$). In addition, 2 more items approached significance ($p = .136, .167$). GLOBE students outperformed non-GLOBE students in demonstrating an understanding of pH in water and water chemistry on these items. For the science inquiry skills scale, GLOBE students scored significantly higher on 2 of the 5 items ($p = .047, .008$). One item

measured students' skill in setting up an investigation of water quality, and the second item required students to interpret changes in satellite images of a region over time.

When all items, regardless of their power to discriminate GLOBE and non-GLOBE students are used in the analysis, there were significant differences across the three samples on hydrology content and inquiry skills scores (see Table 1).

Table 1. Preliminary Analysis of Differences in Scores Across Groups

		Hydrology Content	***Inquiry Skills***
Active GLOBE	Mean	*10.37*	*5.35*
	N	73	57
	SD	3.42	3.03
North Carolina	Mean	*9.69*	*4.04*
	N	13	21
	SD	2.50	1.94
Non-GLOBE	Mean	*7.12*	*3.50*
	N	34	34
	SD	2.91	2.14
		$F = 12.02$ (df = 2, 117), $p < .0005$	$F = 5.84$ (df = 2, 109), $p = .004$

Post-hoc comparisons indicate that on the hydrology content scale, students in North Carolina outscored non-GLOBE students, but their scores were not significantly different from other active GLOBE students' scores. On the inquiry skills scale, North Carolina students' scores were not significantly different from non-GLOBE students' scores, but Active GLOBE students significantly outscored non-GLOBE students. These data should be interpreted with caution, since there was significant attrition in the North Carolina sample from pre-test to post-test. Concerns over the exit examination made finding time to complete the post-test impossible in 3 classrooms there.

Next Steps in the Assessment Development Process

Over the coming year, we expect to further refine our instrument by adding and testing items to our scale to ensure broad coverage of the content and inquiry skills we hope to measure. The need for a large item pool from which to tap student knowledge and understanding, in order to create a valid assessment that is representative of the domain, is an important consideration for future efforts. We will be able to use just under half our items, and more items will be needed to ensure validity of the instrument. In addition, we plan to shift our assessment time frame to the fall, when there are fewer competing pressures on students and concerns about "overtesting" are not as great.

References

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J., Steinberg, L. J., Almond, R. G., Haertel, G., & Penuel, W. R. (2000). *Leverage points for improving educational assessment.* CSE Technical Report. Los Angeles, CA: CRESST.