Chapter 2

Technology Supports for Assessing Science Inquiry

Barbara Means and Geneva Haertel

SRI International

The *National Science Education Standards* (National Research Council [NRC], 1996) place inquiry, applied to scientific content areas, at the core of what it means to be scientifically literate:

Inquiry is central to science learning. When engaging in inquiry, students describe objects and events, ask questions, construct explanations, test those explanations against current scientific knowledge, and communicate their ideas to others. They identify their assumptions, use critical and logical thinking, and consider alternative explanations. In this way, students actively develop their understanding of science by combining scientific knowledge with reasoning and thinking skills. (p. 2)

The *Standards* (NSES) characterize these aspects of science inquiry as a set of "abilities" that all students should exhibit.

Box 2-1 Standards for Science Inquiry, Grades 5-8

- - Identify questions that can be answered through scientific investigations.
- - Design and conduct a scientific investigation.
- - Use appropriate tools and techniques to gather, analyze, and interpret data.
- - Develop descriptions, explanations, predictions, and models using evidence.
- - Think critically and logically to make the relationships between evidence and explanations.
- - Recognize and analyze alternative explanations and predictions.
- - Communicate scientific procedures and explanations.
- - Use mathematics in all aspects of inquiry.

SOURCE: National Research Council, 1996.

This listing of inquiry abilities should not be interpreted as promoting a framework of discrete, linear competencies. The various aspects of inquiry (e.g., "communicate scientific procedures and explanations" and "think critically and logically to make the relationships between evidence and explanations") get flexibly combined with each other in different permutations in the course of different kinds of science investigations (see

Champagne, Kouba, & Hurley, 2000; NRC, 2000). Nor is science inquiry independent of content knowledge.

Page 13

Inquiry includes examining what is already known in order to effectively plan, conduct, and interpret the results of investigations in specific content areas.

THE IMPORTANCE OF ASSESSMENT

It will not be possible to achieve the goal of this kind of scientific literacy for all—or even most—students without the use of assessments of scientific inquiry. Inquiry assessments are needed within classrooms to help teachers diagnose the nature of their students' understanding and to give students feedback about their performance. Science inquiry assessments are needed within the research and evaluation community to make it possible to compare the efficacy of alternative approaches to supporting science learning. Within accountability systems, science inquiry assessments are needed if teachers and systems are to be held accountable for the NRC standards, but more than that, if we are to avoid sending the wrong message to teachers, students, and parents about what it means to "learn science."

Despite the central importance of inquiry, both as a means for students to acquire a deep understanding of science and as a complex set of interrelated knowledge and processes which in and of themselves are targets of instruction, inquiry is the aspect of science that is least likely to be adequately assessed in large-scale accountability systems. Conventional assessment approaches are quite capable of measuring content knowledge and some of the process skills related to science inquiry (e.g., recognizing confounded variables) in a decontextualized manner. They are ill-suited, however, to capturing multifaceted inquiry processes in meaningful contexts. The conduct of complex, hands-on inquiry is missing from most state, national, and international assessments as well as standardized science achievement tests developed by commercial publishers. Instead,

standardized assessments typically emphasize decontextualized factual knowledge (Quellmalz & Haertel, in press). Even when performance or hands-on tasks are administered on a broad scale, their structure and length, and the demand for coverage of a broad range of science content, significantly limit the aspects of inquiry that can be elicited (cf. Baxter & Glaser, 1998).

While classroom assessment practices do not have to conform to the time limits that constrain externally imposed standardized tests, many teachers mimic the format and focus of standardized tests when they are creating assessment tools for classroom use (NRC, 2001a). As a result, even teachers who incorporate extensive inquiry-oriented investigations in their science teaching often score the inquiry work mainly on the basis of "participation" or "completion" and base class grades on conventional tests of factual knowledge from the textbook (Young, Haertel, Ringstaff, & Means, 1998). When classroom assessments do not reflect adequately the engagement required to pursue a line of inquiry or solve a complex problem, the assessment activities are often perceived as dull and disconnected from the hands-on activities (Cognition and Technology Group at Vanderbilt, 1992).

TECHNOLOGY AS CATALYST AND SUPPORT

The NRC report *How People Learn* (1999) makes the point that technology can be used to help teachers understand student thinking and provide meaningful, timely feedback. Nowhere is there greater need and potential for this kind of contribution than in the area of science inquiry

Page 14

(Brophy et al., 2000; Duschl & Gitomer, 1997; White & Frederiksen, 1998). Increasingly, technology plays a major role in science inquiry in all areas of science. If students use tools and data sets with some degree of authenticity when they engage in science investigations, they are using technology. Under such circumstances, it makes sense to think about capitalizing on the data capture capabilities of the technology to preserve student actions for the purposes of assessment.

Over the last decade, technology-based simulations and environments for science inquiry have been a rich area of research and development (with tools such as GenScope, the Knowledge Integration Environment, and ThinkerTools). Because of the importance of feedback in supporting learning, these software environments have incorporated activities with learner feedback that can be considered embedded assessments.

In contrast to standardized tests and the more conventional paper-and-pencil tests used in most science classrooms, these measures of learning embedded in technology-based learning environments reflect the richness and complexity of science inquiry. They provide examples of ways in which learner choices and the explanations developed within the course of inquiry can provide insights into students' thinking without the interruption of a “test-like” series of questions and answers. Multimedia environments offer opportunities to present students with complex, lifelike situations in which they can pursue a sustained investigation or inquiry. Because students can engage in multiple phases of inquiry (for example, planning an investigation into the quality of the water in a given watershed; collecting data within a simulated environment; organizing and analyzing the data they have collected; forming conclusions and communicating their procedure, findings, and explanations), we can tap not just the individual inquiry “abilities” as stipulated in the Standards, but also students' ability to orchestrate these abilities within a complex task. Technology environments have all kinds of capabilities for capturing the process of student inquiry during this sustained investigation (down to the level of the keystroke if we want that much information) and can accommodate the use of a range of approaches and tools, including collaborative problem solving. Table 2-1 summarizes these capabilities and contrasts them with the features of more conventional assessments.

Despite all this potential, in most cases the rich, technology-based inquiry assessments we can point to are so intertwined with the learning systems within which they are embedded as to be impractical for broader administration (Quellmalz, Haertel, Hoadley, Marshall, & Mishook, 2000). That is, they serve their intended assessment function within the system for which they were developed, but they do not solve the problem of how to assess inquiry activities that are not within that particular learning system.

NEED FOR INQUIRY ASSESSMENTS WITH BROADER FOCUS AND SCALE

NRC (2001a) present six different *purposes* for which educational assessments are used: improving learning, informing instruction, grading, placement, promotion, and accountability. We would add research and evaluation—our own focus—as a seventh purpose. As Atkin and colleagues point out, these different purposes involve different types of people making different

Page 15

kinds of decisions, and therefore are best served by different (although, ideally, compatible) kinds of assessments.

We find it useful to augment this classification of assessments according to purpose with two related dimensions—the focus and the intended scale of application of the assessment procedure or instrument. Table 2-1 illustrates our framework. Focus refers to the breadth of student understanding or skill the assessment seeks to capture. Many attempts to get at students' thinking, either within the context of research or within the moment-by-moment assessment practices of teachers, are concerned with a very specific aspect of knowledge or skill—the

Page 16

Recording and Archiving of Responses	• Paper-pencil • Optical scan	• Mechanisms to reveal steps of problem solving (e.g., Internet trace strategies, electronic notebooks for annotations and describing rationale and documentation of steps) • Web pages • Screen shots

• May accumulate responses over time

SOURCE: Adapted from Quellmalz and Haertel, In Press.

content of a single learning activity or even a fragment of an activity. A common example of a broader focus is measurement of understanding and skill at the level of a whole course curriculum. Competence and achievement are broader foci still. These different foci tend to be associated with different purposes (e.g., competence and achievement tend to be associated with accountability systems), but they are logically distinct dimensions, and different combinations do occur. Similarly, the *scale* of the assessment can vary; it can be used for individual diagnosis or for students within a single classroom, within a school or throughout a district or state, or within a given project or program (which could be very small or very large).

One of the things that strikes us in reading the NRC's recent publication *Knowing What Students Know* (2001b) is that, on the one hand, we have large-scale assessment practices that most teachers find wanting for the purpose of informing learning and, on the other hand, we have research-based assessments of very specific aspects of learning. These two types of assessments differ not only in purpose but also in focus and scale. While the research-based assessments provide guidance as to what is important to measure from a learning science perspective and are extremely useful sources of inspiration for new approaches to measurement, they are typically narrow in focus. When critics of embedded assessments and performance assessments deride them as “learning activities,” these misgivings reflect a concern that the assessment is so entwined with one particular instructional activity that it could not be used for broader purposes or on a wider scale. Yet, for the purpose of informing learning within a particular instructional unit, it is all to the good if the assessment is seamlessly intertwined with the instructional content of the learning activity. It is when we want to focus on a broader picture of student understanding and skill, and to do so in classrooms where students have had a range of different learning experiences, that such close coupling becomes problematic.

At SRI's Center for Technology in Learning, we have been working to leverage the capabilities of technology and to adapt ideas from system-specific embedded learning assessments (designed for use with specific modules) to the development of assessments with broader applicability. One important impetus to this work is the need that arises within research and evaluation projects to have measures of learning that tap deeper understanding and inquiry skills, yet do so in a way that provides a "fair test" of learning in a reasonably large sample of classrooms, using a range of different software systems or textbooks. Many of the instructional interventions SRI researchers have studied involve the use of the Internet, and SRI has taken advantage of this infrastructure to develop engaging, complex multimedia assessments for delivery over the Web (Center for Technology in Learning, 1999; Coleman & Penuel, 2000;

Page 17

Means, Penuel, & Quellmalz, 2001; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2000; Quellmalz & Zalles, 1999).

TECHNOLOGY-SUPPORTED PERFORMANCE ASSESSMENTS

We can illustrate this kind of work with one of the assessment tasks we developed for use in evaluating the GLOBE environmental science education program. Students participating in GLOBE follow scientists' protocols and collect environmental data on a local study site. They submit their data to the project database over the Internet and have access to data contributed by 4,000 schools from countries around the world. In recent years, the program has attempted to reinforce aspects designed to promote students' use of the collective GLOBE database to explore questions of their own framing. For our evaluation, we wanted to be able to measure inquiry skills associated with the analysis and interpretation of climate data in both GLOBE and non-GLOBE classrooms.

Box 2-2 Sample Student Justifications for Site Selections

Flagstaff seems like the ideal place for the Winter Olympics to take place. There are about 11 days out of the month of February with sunshine. So the people wanting to watch at the base camp can watch outside with plenty of sunshine and warmth.... The average snowfall for Flagstaff is 1389 mm and it meets the requirements for the O.C. by 389 mm....

Since the temperature at the base level should be at least warm, and if possible, sunny, this proves that Salt Lake City is cool enough compared to Banff, which is too cold -3 degrees Celsius. And, Salt Lake City has up to 5 days of sunshine in February. This keeps the players and spectators more comfortable than if they were at Banff.

Flagstaff best met all of the requirements except in maximum peak temperature. Their temperature was so low, that with the aid of sunlight, their snow could melt.

Elevation, temperature, and the sunny days were all considered when making the choice between the five cities. Although all of the choices would be ideal sites for the winter games only one of the sites could be used. After comparing the data Canada was chosen.

SOURCE: Center for Technology in Learning, 2001.

We have developed several Web-based assessment tasks for our evaluation. One of these tasks, for example, presented students with a set of climate-related criteria for choosing a site for the next Winter Olympics. Given multiple types of climate data on a set of feasible candidate cities, students were asked to analyze the data in terms of the criteria, decide which candi-

date city best met the climate criteria overall, and prepare a persuasive presentation for the Olympic Committee, complete with graphs of relevant climate data contrasting the city they chose with the default candidate (Salt Lake City). From students' performance on this complex task, SRI researchers derived both measures of specific skills, such as the ability to comprehend quantitative information presented in graphic form, and measures of broader aspects of scientific inquiry, such as the ability to communicate and defend a scientific argument (Coleman & Penuel, 2000). The explanations students provided for their choices (see Table 2-2) revealed both

Page 18

confusion concerning certain concepts (e.g., “Their temperature was so low, that with the aid of sunlight, their snow could melt”) and wide variation in the ability to systematically apply a complex set of criteria. Students who identified the objectively “best” city according to the criteria, but did not provide a systematic data-based justification, could be distinguished from students who did both. (Both sets of students would have been similarly successful on a multiple-choice test.) There were also students who did not choose the “best” site but who approached the task systematically and presented an argument and a set of graphs with data consistent with their choice. Table 2-3 presents the scoring scheme for the Olympic task.

While students enjoyed completing the Web-based assessments, and the assessments served the purposes of our evaluation, such assessments, like those embedded within learning systems discussed above, have limitations and do not satisfy broader assessment needs. To date, much of SRI's effort in assessment has been devoted to finding ways to use technology tools to deliver and capture students' performance. As the challenges associated with technology are overcome, we have begun to turn our attention to other limitations of situation-specific,

embedded, Web-based assessment tasks that can impact the validity of the scores they generate and the applicability of the tasks and scores in varying classroom contexts. We note the following limitations of our own and others' work: Each assessment task covers only a narrow piece of curriculum, and a broad set of assessment tasks guided by an assessment framework of inquiry skills within content areas is generally absent. Scoring rubrics are developed for each task or set of tasks used within a given project; their relationship to rubrics used to score other, related tasks used in other projects is not explicit. In some cases, teachers were not involved in the design of the assessment tasks; and the tasks are labor-intensive to develop and score. In light of these limitations, we have concluded that a "one off" approach to assessment will not be sufficient to meet the needs for assessments with a broader focus and scope, as identified in Table 2-4.

At the same time, some of our SRI colleagues have been exploring the use of assessment templates in designing classroom assessment tools. They have implemented this approach to support the GLOBE environmental education program described above. The GLOBE database contains student-collected data from more than 20 protocols in four investigation areas (atmosphere, hydrology, land cover, and soil). SRI has developed templates for assessing students' ability to plan, conduct, analyze, compare, interpret, and communicate investigations with environmental data (Quellmalz, Hinojosa, & Rosenquist, 2001). Teachers have access to a Web-based set of exemplar assessments and to tools for customizing the templates to create their own data inquiry assessments (i.e., they can choose the particular inquiry abilities and type of data with which they want students to work).

Table 2-4 Three Dimensions of Educational Assessment

PURPOSE*	**FOCUS**	**SCOPE OF APPLICATION**
Improving learning	Learning act	Nation
	Instructional module	State

Informing instruction		
Placement	Course	Project/program
Promotion	Competencies or achievement	District
Accountability		School/grade
Research & evaluation		Class
		Individual

SOURCE: Adopted from Natural Research Council, 2001a.

FUTURE DIRECTIONS: PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY

Our experiences developing technology-based assessment tasks for use within evaluation studies and by classroom teachers left us convinced of the potential contributions technology could make to assessment practices, but at the same time highly aware of the need for a more systematic approach to the enterprise. The work of Robert Mislevy and his colleagues (Mislevy,

Steinberg, Breyer, Almond, & Johnson, 1999; Mislevy, Almond, Yan, & Steinberg, 1999; Mislevy et al., 2000) offered a set of principles and a guid-

ing conceptual framework for assessment design, as well as a demonstration that measurement models could be applied to complex assessments such as those needed to assess science inquiry. This "evidence-centered design" framework consists of: (1) a student model, explicating the relationships among the inferences the assessor wants to make about the student; (2) an evidence model, specifying what needs to be observed to provide evidence for those inferences; and (3) a task model, identifying features of the assessment situation that will make it possible for the student to produce that evidence.

Application of the evidence-centered assessment design model and associated statistical techniques has the potential to address many of the issues arising in more situation-specific science inquiry assessment work, such as that performed by SRI, Vanderbilt's Cognition and Technology Group (1992), White and Frederiksen (1998), and Duschl and Gitomer (1997).

Working with Mislevy on an Interagency Educational Research Initiative (IERI) planning grant, we conceived of Principled Assessment Designs for Inquiry (PADI) as an approach to creating assessments for classroom and research use that would cover a broader spectrum of the science curriculum; incorporate cognitive research on learning in specific science domains and in areas of inquiry; build on a robust measurement model; and demonstrate the power of technology to support assessment design, development, delivery, and interpretation.

The essential PADI concept is a system for developing reusable assessment-task templates, organized around schemas of inquiry that are based on research from cognitive psychology and science education. The completed system will have multiple components, including: generally stated rubrics for recognizing and evaluating evidence of inquiry skills; an organized set of assessment development resources; and an initial collection of schemas, exemplar templates, and assessment tasks.

In planning for this project, we quickly realized that if we wanted to develop templates and assessment development tools that would support the work of curriculum developers, we should involve curriculum developers in both the design and the evaluation of the templates and tools. The team for the recently funded PADI implementation project complements SRI's ex-

pertise in science inquiry and technology development and Mislevy's assessment design and psychometric expertise with the science education and curriculum development knowledge of Nancy Songer, principal investigator for the University of Michigan's IERI-funded BioKIDS Project, and Kathy Long, who leads the Full Option Science S ystem (FOSS) project at the Lawrence Hall of Science, University of California, Berkeley. The BioKIDS curriculum consists of eight weeks of inquiry-fostering activities focusing on biodiversity. While the program will be used by tens of thousands of learners nationwide in upcoming years, the primary focus is on 5th and 6th grade students in high-poverty urban classrooms within the Detroit public schools. The BioKIDS curricular sequence includes activities to build students' ownership and control of inquiry thinking over time. Students begin their exploration of biodiversity through focused fall and winter monthly observations of their local schoolyard. Data are collected on animal distribution and seasonal changes across city regions. Students systematically explore data and organize their understandings in the form of species accounts that are compiled in an electronic

Page 21

field guide. Students' own questions focusing on animal distribution, interdependence, and the impact of humans on animal diversity are explored through the comparison of city and national park data on similar species. The PADI assessment will allow BioKIDS to systematically characterize students' understandings over time, as their inquiry understandings develop across the various curriculum units. FOSS middle school courses, each of which requires 9-12 weeks, cover the content areas of earth/space, life, and physical sciences/technology. Lawrence Hall of Science estimates that 60 teachers and 10,000 students have participated in the development and testing of these curriculum units. FOSS focuses on supporting student learning in three areas: understanding science content, conducting investigations, and building explanations. FOSS developers have had great success in developing assessments for the science content and building explanations variables, but have found assessment of the inquiry skills entailed in conducting investigations more of a challenge. The PADI project is ex-

pected to provide a theoretical and practical framework that can advance the FOSS assessment system and provide teachers with critical tools to improve student learning.

In addition to these partnerships with curriculum development projects, the PADI team will be strengthened by the participation of Mark Wilson, professor at the University of California, Berkeley and an expert in the psychometric modeling of cognitive structures. Wilson brings his experience modeling cognitive structures in the area of science inquiry (Roberts, Wilson, & Draney, 1997; Wilson & Sloane, 2000) and his M2RCMI measurement model (Adams, Wilson, & Wang, 1997), which will be used to support the scoring of the assessment tasks.

PADI will have multiple components, including:

- a classification of different types of science inquiry tasks, each of which can become the basis for an assessment "template"
- generally stated rubrics for recognizing and evaluating evidence of inquiry skills within each developed template;
- an organized set of assessment development resources;
- an initial collection of schemas, exemplar templates, and assessment tasks produced in the context of the BioKIDS and FOSS projects; and
- a statistical model that will support rigorous analyses of student learning.

In addition, we will be exploring a multidisciplinary, multi-institutional, co-development process in which knowledge engineers, software developers, psychometricians, content experts, curriculum developers, and teachers form a networked improvement community (NIC) around the design and evaluation of PADI assessment tasks. NIC members will both contribute to and take from the pooled resources of the community.

PROGRESS TO DATE

During the past year's planning grant effort, we applied the PADI conceptual framework to existing assessment tasks from two SRI projects (the

GLOBE classroom assessments described above and a computer-based environment for learning chemistry). Working with the individuals who designed the original assessment tasks, we applied the evidence-centered design

Page 22

framework to produce prototype reusable task templates, built around inquiry schemas in the environmental and physical sciences.

SRI staff who were very familiar with the GLOBE and chemistry curricula but less familiar with the PADI framework, completed the retrofitting process, which involved specifying the student, evidence, and task models for each of the investigation phases included in the science curricula. For the *student model*, we identified those science inquiry concepts and skills that students would be expected to know. In specifying the *evidence model*, we first identified the concepts on which each observable variable would depend. Second, we developed a generalized rubric to score the observations conducted within each investigation phase. The *task model* included the specification of representational forms that student work products would take. An example of a work product is "an ordered list of free-form phrases describing the steps in an investigation plan." The task model also included the presentation materials and properties that students would use in creating their work products (e.g., tools, technology affordances, and materials). For example, students might be asked to create a drawing or animation that illustrates a phenomenon or to record data in a log. By retrofitting these assessment tasks to the PADI framework, we were able to demonstrate the usability of the PADI design processes with individuals who were new to the approach, but whose skills and backgrounds were similar to those of our curriculum development partners.

RESEARCH ON QUALITY OF EVIDENCE OF STUDENT LEARNING AND SCALABILITY

The PADI project will conduct research on whether the assessments that are generated provide better evidence about students' inquiry skills and

whether the PADI design process is scalable. Working with our curriculum development partners, we will conduct an evaluative study to examine the quality of evidence yielded by the PADI assessments. We will compare the evidence of student inquiry from three sources: cognitive analyses (think-alouds) of inquiry problems, inquiry tasks used as part of large-scale reference exams (e.g., NAEP, TIMSS, or New Standards), and the newly developed PADI assessment tasks.

To better understand the scalability of the PADI process, we will study the assessment design and implementation process. To achieve scalability, we seek to develop our conceptual framework, implementation framework, templates, and design supports at a level of generality that can be applied in different science content areas. The FOSS and BioKIDS implementation sites will provide access to hundreds of middle school students and teachers in diverse settings. The contributions of FOSS and BioKIDS curriculum developers and teachers, and the problems they encounter with our tools and assessments, will be documented in a qualitative study. We will also describe the use of the assessments in different classroom contexts, including urban schools and schools with considerable experience implementing inquiry science, as well as those with less experience.

Page 23

THE ROLE OF TECHNOLOGY IN PADI

Technology will support almost every component of PADI. The various categories of science inquiry tasks will be realized as reusable software templates that allow curriculum developers to “fill in the slots” in much the way GLOBE teachers customize the classroom assessments described above. SRI staff will work with BioKIDS and FOSS to develop Web-based exemplar assessment tasks using these templates. A software instantiation of Wilson's M2RCMI measurement model will be applied to the tasks. Electronic communication and online repositories of resources will support the networked improvement community (NIC). Thus, technology will play an important role in the design, dissemination, presentation, and scoring of PADI assessment tasks.

CONCLUSION

The recent National Research Council publication *Knowing What Students Know* (2001b) asserts that "Developers of educational curricula and classroom assessments should create tools that will enable teachers to implement high-quality instructional and assessment practices, consistent with modern understanding of how students learn and how such learning can be measured" (p. 306). In complex domains, such as science inquiry, where the knowledge and skills being assessed are numerous, interdependent, and executed over an extended timeframe, it is unlikely that this goal can be attained without the use of technology supports.

REFERENCES

Adams, R.J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement* , *21*(1), 1-23 .

Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessment. *Educational Measurement: Issues and Practices* , *17*(3), 37-45 .

Brophy, S., Elder, S., Pfaffman, J., Martin, T., Mayfield, C., Vye, N., & Zech, L. (2000). Expanding new methods of technology-embedded assessment and instruction. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Center for Technology in Learning. (1999). *GLOBE year 4 evaluation: Evolving implementation practices* . Menlo Park, CA : SRI International .

Center for Technology in Learning. (2001). *GLOBE year 5 evaluation* . Menlo Park, CA : SRI International .

Champagne, A.B., Kouba, V.L., & Hurley, M. (2000). Assessing inquiry. In J. Minstrell & E. van Zee (Eds.), Inquiring into inquiry teaching in science. Washington, DC : American Association for the Advancement of Science .

Cognition and Technology Group at Vanderbilt. (1992). The Jasper series as an example of anchored instruction: Theory, program description, and

assessment data. *Educational Psychologist* , 27 , 291-315 .

Page 24

Coleman, E., & Penuel, W.R. (2000). Web-based student assessment for program evaluation. *Journal of Science Education and Technology* , 9 , 327-342 .

Duschl, R.A., & Gitomer, D.H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment* , 4 , 37-73 .

Means, B., Penuel, W., & Quellmalz, E. (2001). Developing assessments for tomorrow's classrooms. In W. Heinecke & L. Blasi (Eds.), Research Methods for Educational Technology. Volume One: Methods of Evaluating Educational Technology. Greenwich, CT : Information Age Press .

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Ba yes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H. Prade (Eds.), Proceedings of the fifteenth conference on uncertainty in artificial intelligence (pp. 437-446). San Francisco : Morgan Kaufmann .

Mislevy, R.J., Steinberg, L.S., Almond, R.G., Haertel, G.D., & Penuel, W.R. (2000, February). Leverage points for improving educational assessment. Paper prepared for the Technology Design Workshop, SRI International, Menlo Park, CA.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999, April). A cognitive task analysis, with implications for designing a simulation-based performance assessment. Presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

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

National Research Council. (1999). *How people learn: Brain, mind, experience, and school* . Committee on Developments in the Science of Learning.

J.D. Bransford, A.L. Brown, & R.R. Cocking (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC : National Academy Press .

National Research Council. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning* . Washington, DC : National Academy Press .

National Research Council. (2001a). *Classroom assessment and the National Science Education Standards* . Atkin, M.J., Black, P., & Coffey, J. (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC : National Academ y Press .

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment* . Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC : National Academ y Press .

Quellmalz, E., & Haertel, G. (in press). Breaking the mold: Technology-based science assessment in the 21st century.

Quellmalz, E., Haertel, G.D., Hoadley, C., Marshall, S., & Mishook, J. (2000). *21st century assessment planning grant: Final report* (PDU-99-086). Menlo Park, CA : SRI International .

Quellmalz, E., Hinojosa, T., & Rosenquist, A. (2001). Design of student assessment tools for the Global Learning and Observations to Benefit the Environment (GLOBE) program. Presentation at the annual GLOBE International Conference, Blaine, WA.

Quellmalz, E., & Zalles, D. (1999). *World student assessment report: 1998-99* . Menlo Park, CA : SRI International .

Roberts, L., Wilson, M., & Draney, K. (1997, June). *The SEPUP assessment system: An overview (BEAR Report Series, SA-97-1)* . University of California , Berkeley .

White, B.Y., & Frederiksen, J.R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction* , 16 , 3-117 .

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education* , 13(2), 181-208 .

Young, V.M., Haertel, G., Ringstaff, C., & Means, B. (1998). *Evaluating global lab curriculum: Impacts and issues of implementing a project-based science curriculum* . Menlo Park, CA : SRI International .