Reconsidering Large-Scale Assessment to Heighten Its Relevance to Learning

Lorrie A. Shepard

Lorrie Shepard

Many science teachers have been affected indirectly by high-stakes, accountability pressures as they watch attention and resources flow to language arts and mathematics instruction—because these subjects are tested. Others have experienced firsthand the ways that external science assessments can undermine inquiry-based curricula and efforts to teach for understanding. Is it possible to counteract these effects and make external, large-scale assessments more relevant to student learning? How can large-scale assessments, remote from the classroom, serve instructional purposes?

I agreed to write a chapter addressing these questions with some trepidation because the history of assessment reform has not been pretty. Ideally, evaluation data should be used to improve instructional programs and thus ensure meaningful learning opportunities for students. The difficulty with promoting an ideal, however, is that we have all seen how a lofty goal can be corrupted when pursued on the cheap or when too many participants hold conflicting ideas about what was intended. A decade ago, standards-based reformers, recognizing the deleterious effects of traditional, multiple-choice tests on ambitious learning goals, promised to create "authentic assessments" and "tests worth teaching to." These promises have not been realized, however, in part because accountability advocates have pursued the slogan of high standards without necessarily subscribing to the underlying theory calling for profound changes in curriculum, instruction, and assessment.

The central aim of this chapter is to consider how large-scale assessments could be redesigned to heighten their contribution to student learning. In this section, which acts as a preamble, I (1) explain why assessments must be designed and validated differently for different purposes and the implications of this differentiation for large-scale assessments and (2) summarize the essential features of effective classroom assessment. While classroom assessment is not the focus of this chapter, we cannot consider here how large-scale assessment could be made compatible with and supportive of classroom instruction and assessment without a shared understanding of effective classroom assessment. In the next, main section of the chapter I address the important purposes served by large-scale assessment: (1) exemplification of learning goals, (2) program "diagnosis," and (3) certification or "screening" of individual student achievement. In addition, large-scale assessments can serve as a site or impetus for professional development to enhance the use of learning-centered classroom assessment. I conclude with an analysis of the impediments to change and recommendations for addressing these challenges.

Assessments Designed for Different Purposes

To the layperson, a test is a test. So why couldn't the same test be used to diagnose student learning needs; to judge the effectiveness of teachers, schools, districts, and states; and to compare U.S. schools to the schools of other nations? For measurement specialists, however, purpose matters. Purpose shapes test design and alters the criteria for evaluating the reliability and validity of the test. According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME 1999), "No test will serve all purposes equally well. Choices in test development and evaluation that enhance validity for one purpose may diminish validity for other purposes" (145).

Large-scale assessments are used to monitor achievement trends over time and to hold schools and school districts accountable. In some states, large-scale assessments are also used to make high-stakes decisions about individual teachers and students—for example, in regard to teacher pay increases, grade-to-grade promotion, or graduation from high school. Because of the significant consequences that follow from the results, large-scale assessments must be highly reliable. Thus, purpose shapes technical requirements. And, to be fair, large-scale assessment data must be collected in a standardized way to ensure comparability across schools. It would be unfair, for example, if one school gave the test a month later than other schools, explained unfamiliar words to students, or allowed extra time when students hadn't finished.[1] Because of the cost of ensuring reliability and standardization and because of the intrusion on instructional time, large-scale assessments are administered only once per year and must necessarily be broad, "survey" instruments touching lightly on the many curricular topics and skills taught throughout the year.

In contrast, classroom assessments intended to help students learn must be closely

- engages students in self-monitoring of their own learning,
- makes the features of good work understandable and accessible to students, and
- provides feedback specifically targeted toward improvement.

These elements can be made a part of everyday instructional routines, using a definition of formative assessment developed by Sadler (1989) and another recent National Research Council report, *Classroom Assessment and the National Science Education Standards* (Atkin, Black, and Coffey 2001). For assessment to be *formative* in the sense of moving learning forward, three questions are asked: (1) Where are you trying to go? (2) Where are you now? (3) How can you get there? It is because of the explicitness of these steps and the focused effort to close the gap between 1 and 2 that assessment actually contributes to learning. Elsewhere I have also argued that effective use of these strategies requires a cultural shift in classrooms so that students are less concerned about grades and hiding what they don't know and are more focused on using feedback and support from teachers and classmates to learn—that is, to solve a problem, improve a piece of writing, or figure out *why* an answer is correct.

Finally, to be effective, classroom assessment will need to find ways to address the many negative effects of grading on student motivation. Cognitive studies have shown us that making criteria explicit will improve student-learning outcomes (Fredericksen and Collins 1989). But motivational psychologists have found that traditional grading practices may negatively affect students' intrinsic motivation, their sense of self-efficacy, and their willingness to expend effort or tackle difficult problems. Therefore, merely sharing grading criteria will not automatically eliminate the negative effects of grading.

Unlike the extensive amount of work on formative assessment in recent years, there has been much less attention, outside of the motivational literature, to the type of grading policies that would improve rather than decrease motivation. Self-assessment is one example of a change in classroom practice that could serve both cognitive and motivational ends. Self-assessment makes the features of excellent work explicit and helps students internalize these criteria (thus serving cognitive purposes). At the same time, asking students to self-assess according to well-defined criteria establishes a mastery rather than normative definition of success, conveys developing competence, and illustrates how effort could lead to improvement, all of which enhance motivation (Stipek 1996). More work needs to be done to relate formative and summative assessment within classrooms. Perhaps all formative assessment should be reserved exclusively for learning purposes, not for grading—even while eventual summative criteria are used formatively. Note that pursuit of this idea would run against the highly litigious point systems that many teachers currently use to track every assignment and to justify grades.

Purposes Served By Large-Scale Assessment

Large-scale assessments such as the Third International Math and Science Survey (TIMSS), the National Assessment of Education Progress (NAEP), and various state- and district-level assessment programs are used to measure student achievement for aggregate units (nations, states, districts, schools), to track changes in achievement for these units over time, and sometimes to measure the performance of individual students. If the content of a large-scale assessment adequately represents ambitious curricular goals—as called for in the science standards, for example—then large-scale assessment can become an integral part of curricular reform and instructional improvement efforts. Such an assessment program could be used to: exemplify important learning goals; diagnose program strengths and weaknesses; report on the proficiency status of individual students; and, through associated professional development opportunities, improve teachers' abilities to teach to the standards and at the same time become more adept in using formative assessment. These purposes would not be served, however, by traditional, multiple-choice-only tests that do not adequately embody the National Science Education Standards.

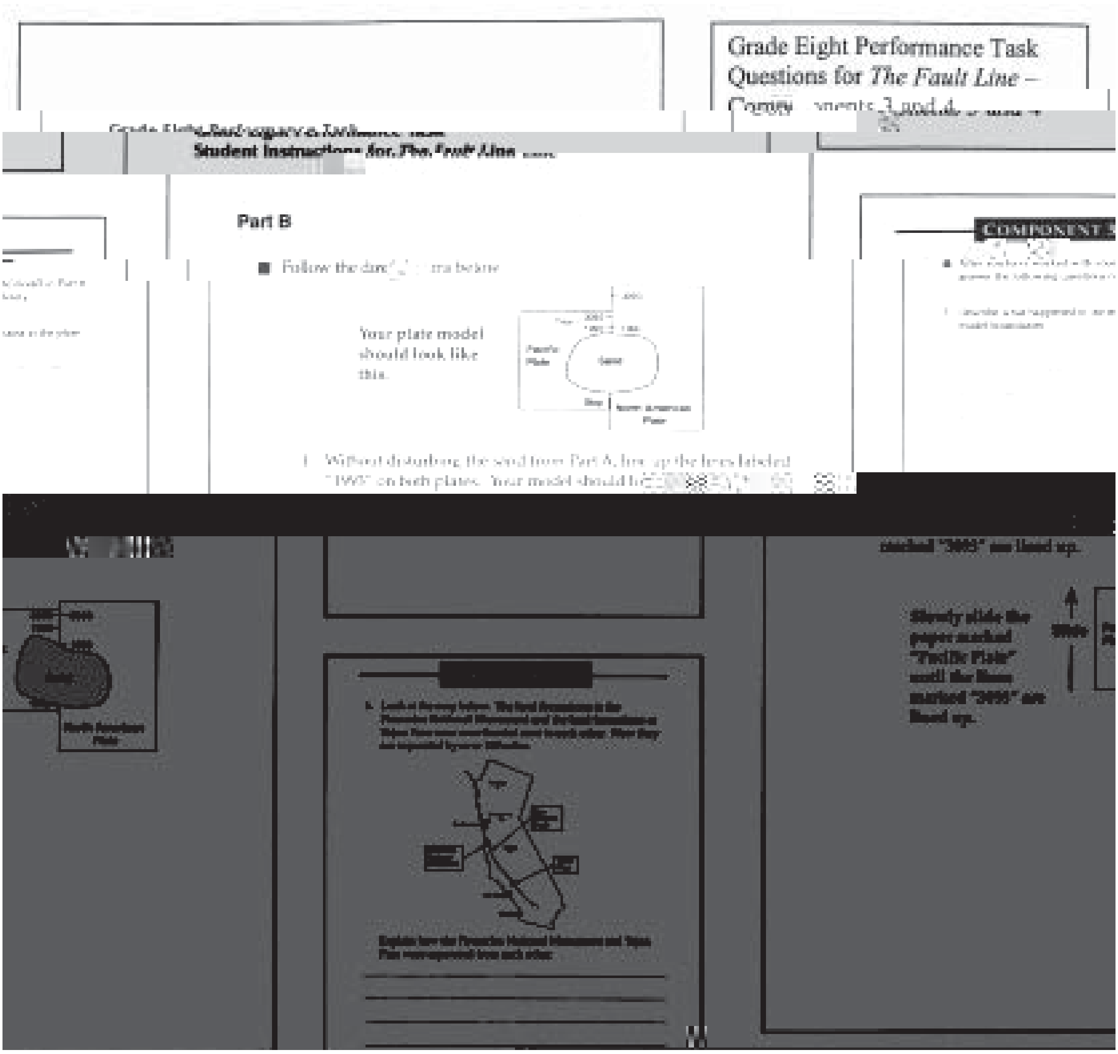
Exemplification of Learning Goals

The science standards developed a vision for science instruction by drawing on best practices, but for many teachers the standards call for significant changes in practice—away from vocabulary-laden textbooks and toward more inquiry-based approaches. For many, these hoped-for changes may seem out of reach either conceptually or practically. Large-scale assessments can give life to the standards expectations by illustrating the kinds of skills and conceptual understandings that students are expected to have mastered. Moreover, because some of the very best assessment tasks would also qualify as good instructional activities, released assessment items can help to raise awareness about the kinds of instructional opportunities students need if they are to develop deep understandings and effective inquiry skills.

The performance task illustrated in Figure 1 is taken from *A Sampler of Science Assessment* developed by Kathy Comfort and others in the California Department of Education (1994). The task gives eighth-grade students hands-on experience with subduction and asks them to generalize their understandings from the physical model to information about California landmarks. One could reasonably expect that students who had had previous instruction on geological processes and plate tectonics would do well on this task. If, however, students with textbook exposure to these ideas faltered in providing explanations, the assessment experience might prompt teachers to consider using more conceptual learning tools in the future, and, in fact, the investigation shown in Figure 1 is an example of the type of instructional activity needed.

Figure 1.

(Figure 1. continued)



In some cases a single conceptual question, if used reflectively, can prompt teachers to reconsider the efficacy of their instructional approach. In some sense, Phil Sadler's classic films, *A Private Universe* and *Minds of Our Own* are each based on one significant conceptual question. Can you explain what makes the seasons? Can you use a wire, a bulb, and a battery and make the bulb light? The fact that so many Harvard graduates struggled with the first question, and MIT graduates with the second, has prompted many science teachers to think again about what their students are really understanding when they pass traditional tests. Thus, if a state assessment reflects the National Science Education Standards it serves both as a model of what's expected for student mastery and also of the kinds of instructional activities that would enable that mastery.

In preparing to write this chapter, I asked experts in several states to comment on my outline of large-scale assessment purposes and to provide examples of each application where appropriate. Rachel Wood, Education Associate in Science, and Julie Schmidt, Director of Science, are part of the science leadership team responsible for the development of Delaware's Comprehensive Assessment Program. They responded with a detailed commentary, recounting their experiences in involving science teachers in development of summative assessments for curriculum modules (as part of the know whether its students were doing relatively better (compared to state normative data) on declarative knowledge items or on problems requiring conceptual understanding. It is also possible to report on relative strengths and weaknesses according to content categories: life science, physical science, Earth and space science, science and technology, and science in personal and social perspectives. This type of profile analysis would let a school or district know whether its performance in technology was falling behind performance in other areas, or whether there were significant gender effects by content category. For example, we might anticipate that girls would do better in science programs that emphasize the relevance of science to personal and social perspectives, while boys might do relatively better in applications of technology. Results such as these might prompt important instructional conversations about how to teach to strengths while not presuming that either group was incapable of mastering material in their traditional area of weakness.

In addition to subtest profiles, particular assessment items can sometimes yield important program diagnostic information. Wood and Schmidt (2002) provide the following examples of conceptual errors and skill weaknesses revealed by assessment results that warranted attention in subsequent professional development efforts.

> *For instance, an eighth-grade weather assessment revealed that students across the state have over-generalized their knowledge of the movement of all air masses as having to go from west to east. In the classroom, students are studying the movement of weather fronts and predicting weather patterns, many of which do move from west to east. That piece of understanding has now been applied to the movement of all air masses. They are unable to explain ocean breezes on the east coast with this model or Bermuda highs that they experience in their daily lives. This information was not uncovered through a question about weather patterns in the United States but by using a question on land and sea breezes. There is now an opportunity to address this issue in professional development because this suggests that the idea originates from some connection made in the classroom. This confirms what we mentioned earlier, that students are indeed constructing knowledge in the classroom that teachers might not be aware of unless they search for it. Most teachers are probably delighted that students have the idea that most weather fronts move from west to east, but were unaware that students would over-generalize, unless the class has an opportunity to work through the limits of a "rule" or model.*

And a second example:

> *Analysis of item statistics from the state test reveals major weaknesses that the leadership can address through professional development. For example, questions asking students to construct or interpret a simple graph indicate*

that students were not being given enough opportunities to graph data and analyze the results, compare graphs, or draw conclusions from the kind of graph that might appear in the newspaper, etc.… One item, for example, with a P-value of .31 in simple graphing indicated an alarming weakness. A P-value of .80 was expected. As a result the leadership selected graphing items, rubrics, and samples of student responses with P-values to focus discussion on the instructional implications of the student responses…. Some of the lead teachers participated in the piloting of released items and were stunned that their own students were performing at a level that confirmed the P-value found for the whole state.

Because large-scale assessments are broad survey instruments, test makers often have difficulty providing very detailed feedback for curriculum evaluation beyond major subtest categories. This is especially true for assessments like TIMSS and NAEP that cross many jurisdictions and may also be true for state assessments when each district has its own curriculum. Cross-jurisdictional assessments invariably become more generic in how they assess achievement, using questions that call for reasoning with basic content (like on the ACT) rather than presenting the type of question that would be appropriate in a specific course examination. The need for items to be accessible to all students, regardless of what particular science curriculum they have followed, explains why so many NAEP items, for example, involve reading data from a table or graph to draw an inference or support a conclusion, because such items are self-contained and do not presume particular content knowledge. Unfortunately, generic, reasoning items are not very diagnostic nor do they further the goal of exemplifying standards.

How then could we have more instructionally relevant items, like the earlier California example? If state assessment frameworks were to stipulate specific in-depth problem types they intended to use, there would a danger that teachers would teach to the specific item types instead of the larger domain. Conversely, if different in-depth problems were used each year representing the full domain, teachers would be likely to complain about the unpredictability and unfairness of the assessment. Again, I quote extensively from commentary by Wood and Schmidt (2002). They have documented the power of released items (accompanied by student papers and scoring guides) both to exemplify standards and to diagnose gaps in students' learning. Here's how they wrestled with the dilemma of fostering teaching to standards without encouraging teaching to the test.

Many classroom teachers who haven't had the opportunity to be directly engaged in the lead teacher program hold a different view of the test and items than those involved in the assessment development. For instance, classroom teachers express frustration at the comprehensive nature of the standards and not being able to determine "what items" are going to be on *the next test. They complain that we don't release entire forms each year for their students to practice in the classroom in preparation for the next year's test. What has been released and is preferable to release are not isolated items matched to a standard, but an insightful commentary about how and where the concepts in the released items fit into a larger sequence of student conceptual understanding. Teachers will revert back to second-guessing the test items if presented a released item decontextualized from an analysis that helps explain how and why students are struggling with the concept that the item is measuring. For example, when many high school students were unable to construct a simple monohybrid Punnett square and determine the genotypes of both parents and offspring, teachers could easily have thought, "I taught them that, they should know it" or "I guess I need to teach more Punnett squares"—which suggests that it is being taught in a mechanical approach. But the commentary around the released item attempts to turn teachers' attention toward thinking about how students have acquired only a mechanical sense and don't understand why you would have a Punnett square in the first place.*

The example in Figure 2 shows how the analysis accompanying the released item is intended to focus attention on underlying concepts that students might not be understanding. "This particular item taps both procedural and conceptual knowledge, while most teachers think it is only procedural knowledge" (Wood and Schmidt 2002). Because teachers focus on procedural knowledge, students assume the Punnett square is an end in itself rather than a tool for reasoning through possible gene combinations. Lacking conceptual knowledge, they are likely to stack up illogical numbers of alleles in each cell. Wood and Schmidt's analysis is intended to try and reconnect the specific test question to a larger instructional domain, which should be the appropriate target of improvement efforts.

Certification or "Screening" of Individual Student Achievement

Historically, many state assessment programs were designed to imitate the NAEP; they provided broad content coverage and were used primarily for program evaluation. NAEP does not produce individual student scores. In fact, using the strategy of matrix sampling, each participating student takes only a small fraction of the items in the total test pool so as to minimize testing time and ensure a rich representation of the content domain. In recent years, under pressure to provide more accountability information, many assessment programs have abandoned their matrix sampling designs and instead give the same test to every student so that individual scores can be reported. The No Child Left Behind Act requires all states to produce student scores in reading and mathematics in grades three to eight, with testing in science in certain grades to begin in 2007–2008. Individual reporting of students' proficiency status is a type of certification testing, not unlike a licensure test—with accompanying

Figure 2.

LIFE SCIENCE

Analysis:

Released Item:

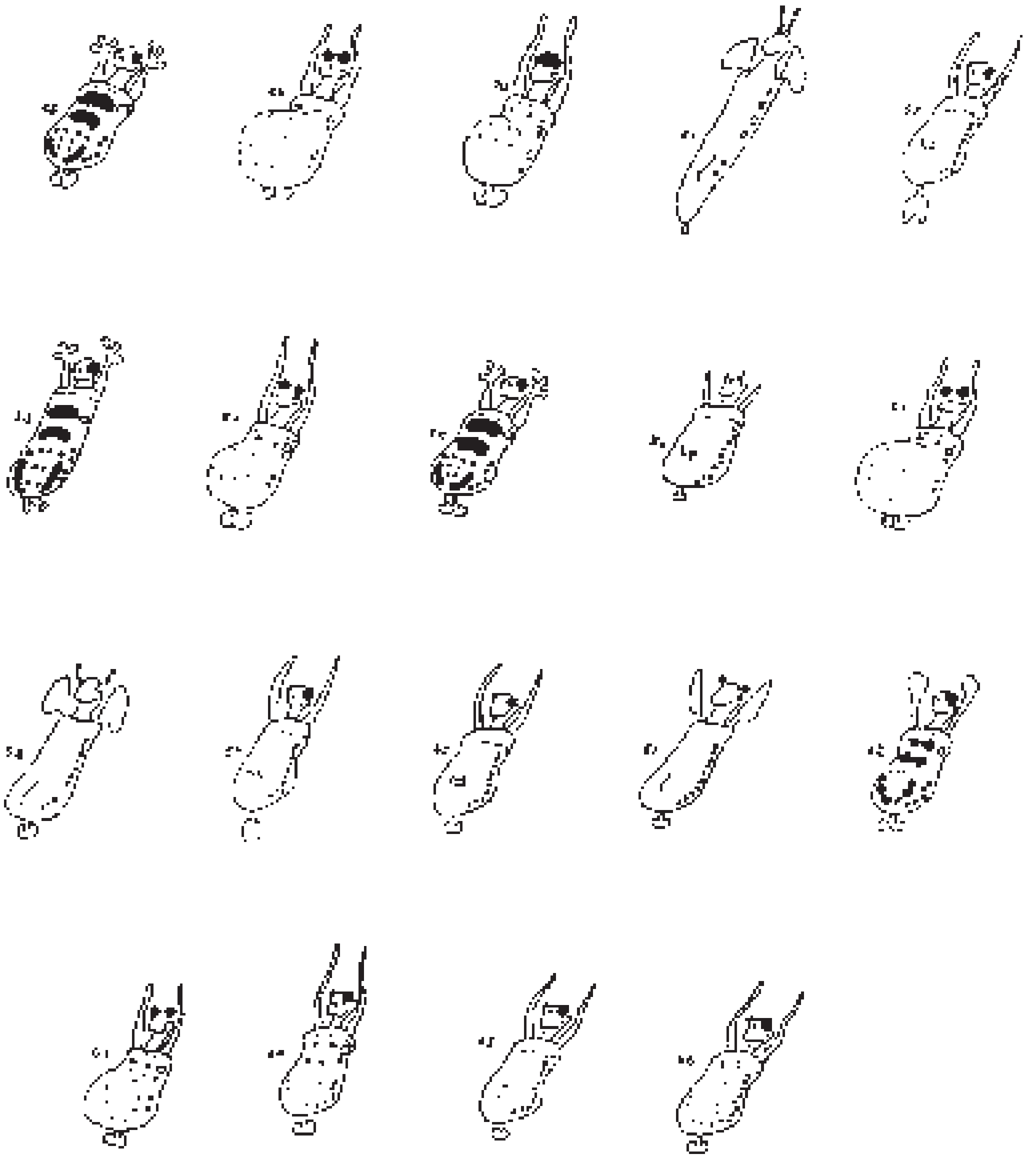
Scoring Tool:

requirements for technical accuracy. When used for high-stakes purposes, tests must be designed with sufficient reliability to yield a stable total score for each student. This means that, within a reasonably small margin of error, students would end up in the same proficiency category if they were retested on the same or closely parallel test.

Reliability does not ensure validity, however. Especially, reliability cannot make up for what's left out of the test or how performance levels might shift if students were allowed to work with familiar hands-on materials, to work in groups, to consult resources, or to engage in any other activities that sharply changed the context of

[illegible]	#	[illegible]
[illegible]	[illegible]	2 [illegible] 20
[illegible]	[illegible]	20 [illegible] 1
[illegible]	[illegible]0 [illegible]	1 [illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible] 2	[illegible]

(Continues on next page.)



for teachers to use to check on student progress but not to be used in formal data collection.

Professional Development

Professional development associated with standards-based reforms has tended to focus on the intention of the standards (Why should students be able to communicate mathematically?), and on curriculum materials and instructional strategies to implement the standards (What does inquiry-based instruction look like?). Assessment activities tied to standards have the potential to deepen teachers' understandings of the meaning of standards as well as to provide the means to improve student learning. The additional goal of having teachers become more adept at using specific formative assessment strategies can also be furthered by professional development that addresses content standards. There are two important reasons for embedding teachers' learning about assessment in larger professional development efforts—one practical, the other conceptual. First, teachers' time is already overburdened. It is very unlikely that teachers could take time to learn about formative assessment strategies in a way that is not directly tied to the immediate pressures to raise student achievement on accountability tests. Second, assessment efforts only make sense if they are intimately tied to content learning. Therefore, assessment learning can be undertaken in the context of helping teachers improve performance on a state test, so long as we clearly understand the difference between teaching to the standards and teaching the test.

Folklore of advanced placement (AP) examinations has it that some teachers return to Princeton year after year to participate in the scoring of AP exams because of the learning experience. Not only is it important to see what kinds of questions are asked, but it makes one a better teacher to engage with student work and to discuss with one's colleagues how to interpret criteria in light of specific student performances. In this same vein, Wood and Schmidt (2002) describe several different aspects of professional development that occurred in Delaware when teachers were involved in assessment development, pilot testing, and scoring. First and foremost, "teachers became hooked on student learning." By focusing on what students were learning, they moved from being good at delivering inquiry-based instruction to focusing on what students were actually learning from that instruction. For example, teachers learned to use double-digit rubrics that produced both a score and the reason for the score, "which completely transformed our thinking."

> *A single-digit rubric just lumps partially correct responses together and doesn't discriminate between the milder and more serious partially correct or wrong responses. The diagnostic rubrics are ordered so that teachers score student work and easily flag the most frequent missteps in student thinking. This kind of diagnostic information is not available from a single-digit rubric that is so holistic that it fails to identify that students get things*

> *their confidence in knowing what their students know. It was Shavelson (also a consultant to the project) who encouraged the leadership to let teachers struggle through this new "problem space" because ultimately that is where all learning occurs. An opportunity to discuss not only their students' learning but similarly situated students' learning with other teachers using the same units has proven to be a key ingredient for realizing Fullan's idea of assessment conversations and is a more powerful mode of professional development than learning the modules and inquiry-based teaching without this aspect.* (Wood and Schmidt 2002)

To summarize, then, professional development focused on assessment of student learning can be a powerful tool to help teachers move beyond merely implementing inquiry activities to an increased awareness of what their students are getting from the activities. Given the layers of assessment-related demands already faced by teachers, efforts to improve classroom assessment strategies should be woven into standards-based professional development and curriculum development. Teachers need better access to materials that model teaching for understanding—with extended instructional activities, formative assessment tasks, scoring rubrics, and summative

Finally, there is the difficulty that policy makers may hold very different beliefs about standards-based reform than those who originally advocated for conceptually linked curriculum and assessment reforms. While originators like Smith and O'Day (1990) and Resnick and Resnick (1992) were clear about the need for what they called *capacity building*, including substantial professional development for teachers, many present-day policy makers have adopted an economic incentives model as their underlying theory of the reform. Those holding the latter view are unlikely to see the need to invest in curriculum development or professional training. Add to this picture the fact that "data-driven instruction" is being marketed more aggressively than are rich assessment and curriculum units. Using data to guide instruction is, of course, a good thing. Investing in mechanical data systems is a mistake, however, if they are built on bad tests. There is no point in getting detailed disaggregations of test data when test content bears little resemblance to valued curriculum. Trying to make sense of this cacophonous scene will be difficult. What one should advocate for will clearly be different in each state depending on the quality of the existing large-scale assessment and likelihood of persuading state-level decision makers to invest in instructionally relevant curriculum development and professional training.

If science educators want to move toward large-scale assessment that is conceptually linked to classroom learning, what should they be *for?* They should advocate for a good test that embodies the skills and conceptual understandings called for in the science standards. A rich and challenging assessment could take the form of curriculum-embedded assessments or be a combination of state-level, on-demand assessments and local embedded assessments, projects, and portfolios as in the New Standards Project (1997). As advocated in *Knowing What Students Know* (Pellegrino, Chudowsky, and Glaser 2001), there should be a strong substantive coherence between what is called for in the state assessment and what is elaborated in local instructional units and classroom assessments. To realize the full potential for teacher learning, professional development should be provided that uses the power of assessment to look at student work and to redesign instruction accordingly. Teachers should have access to curriculum materials that reflect inquiry-based instruction with well-conceived assessment tools built in. And they should have supported opportunities to try out new instructional materials and formative assessment strategies.

What if the state has a bad test? Then the strategies for science educators should be quite different. In fact, the goal should be to reinvigorate the intended goals for learning and to be explicit about what would be left out if we focused narrowly on the curriculum implied by the test. Groups of teachers or curriculum specialists might want to go through this exercise of mapping the state test to the science standards. Then they could ask, What support is needed to ensure that instruction focuses on the standards rather than the test, and what evidence will we provide to parents and school board members to educate them about important accomplishments not reflected in the test?

Ultimately the goal of any assessment should be to further student learning. Classroom assessments have the greatest potential for directly improving learning because they can be located in the midst of instruction and can provide timely feedback at just the point of a student's uncertainty or incomplete mastery. Large-scale assessments can also support the learning process, but to do this they must faithfully elicit the knowledge, skills, and reasoning abilities that we hope for students to develop, and they must be linked in a well-articulated way to ongoing program evaluation and professional development.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). 1999. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Atkin, J. M., P. Black, and J. Coffey. 2001. *Classroom assessment and the national science education standards*. Washington, DC: National Academy Press.

Black, P., and D. Wiliam.1998. Assessment and classroom learning. *Assessment in Education* 5(1): 7–74.

California Department of Education. 1994. *A sampler of science assessment*. Sacramento: California Department of Education.

Fredericksen, J. R., and A. Collins. 1989. A systems approach to educational testing. *Educational*