

SCIENCE EDUCATORS' ESSAY COLLECTION

Everyday Assessment



in the Science Classroom

Edited by J Myron Atkin and Janet E. Coffey

NSTApress®

NATIONAL SCIENCE TEACHERS ASSOCIATION

Arlington, Virginia



Claire Reinburg, Director
J. Andrew Cocke, Associate Editor
Judy Cusick, Associate Editor
Betty Smith, Associate Editor

ART AND DESIGN Linda Olliver, Director
NSTA WEB Tim Weber, Webmaster
PERIODICALS PUBLISHING Shelley Carey, Director
PRINTING AND PRODUCTION Catherine Lorrain-Hale, Director
 Nguyet Tran, Assistant Production Manager
 Jack Parker, Desktop Publishing Specialist
PUBLICATIONS OPERATIONS Erin Miller, Manager
sciLINKS Tyson Brown, Manager
 David Anderson, Web and Development Coordinator

NATIONAL SCIENCE TEACHERS ASSOCIATION
Gerald F. Wheeler, Executive Director
David Beacom, Publisher

Copyright © 2003 by the National Science Teachers Association. Chapter 2, "Learning through Assessment: Assessment for Learning in the Science Classroom," copyright © 2003 by the National Science Teachers Association and Anne Davies.
All rights reserved. Printed in the United States of America by Victor Graphics, Inc.

Science Educators' Essay Collection
Everyday Assessment in the Science Classroom
NSTA Stock Number: PB172X

05 04 03 4 3 2

Library of Congress Cataloging-in-Publication Data

Everyday assessment in the science classroom / J. Myron Atkin and Janet E. Coffey, editors.

p. cm.—(Science educators' essay collection)

Includes bibliographical references and index.

ISBN 0-87355-217-2

1. Science—Study and teaching (Elementary)—Evaluation. 2. Science—Study and teaching (Secondary)—Evaluation. 3. Science—Ability testing. I. Atkin, J. Myron. II. Coffey, Janet E. III. National Science Teachers Association. IV. Series.

LB1585.E97 2003

507'.1—dc21

2003000907

NSTA is committed to publishing quality materials that promote the best in inquiry-based science education. However, conditions of actual use may vary and the safety procedures and practices described in this book are intended to serve only as a guide. Additional precautionary measures may be required. NSTA and the author(s) do not warrant or represent that the procedures and practices in this book meet any safety code or standard or federal, state, or local regulations. NSTA and the author(s) disclaim any liability for personal injury or damage to property arising out of or relating to the use of this book including any of the recommendations, instructions, or materials contained therein.

Permission is granted in advance for reproduction for purpose of classroom or workshop instruction. To request permission for other uses, send specific requests to: **NSTA Press**, 1840 Wilson Boulevard, Arlington, Virginia 22201-3000. Website: www.nsta.org

Reconsidering Large-Scale Assessment to Heighten Its Relevance to Learning

Lorrie A. Shepard

Lorrie Shepard is dean of the School of Education at the University of Colorado at Boulder. She has served as president of the American Educational Research Association, president of the National Council on Measurement in Education, and vice president of the National Academy of Education. Her research focuses on psychometrics and the use and misuse of tests in educational settings. Specific studies address standard setting, the influence of tests on instruction, teacher testing, identification of mild handicaps, and early childhood assessment. Currently, her work focuses on the use of classroom assessment to support teaching and learning.

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 learn-

ing 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 tied to particular units of instruction and must be used in the particular days and weeks when students are learning specific concepts. To be truly diagnostic, teacher's questions must probe students' understandings and push to identify extensions where mastery is incomplete or where misconceptions impede learning. Because formative assessment in classrooms is intended to help target the next instructional steps, not to assign official proficiency status, there is much less need for formal assessment procedures or adherence to strict reliability standards. Mismeasurement of a student's knowledge and skills by a teacher one day can be overturned by subsequent assessments in the next day or week.

¹Note that sources of unfairness to students caused by differences in students' experiences with and opportunities to learn tested content are not corrected by standardization.

Knowing What Students Know (Pellegrino, Chudowsky, and Glaser 2001) is a landmark report published recently by the National Academy of Sciences that brings together the knowledge bases of cognitive learning theory and measurement science. Its authors similarly address this link between purpose and assessment requirements, referring to the inevitability of *trade-offs* in assessment design: “Ironically, the questions that are of most use to the state officer are of the least use to the teacher” (224). One way to help policy makers understand the limitations of an external, once-per-year test for instructional purposes is to point out that good teachers should already know so much about their students that they could fill out the test booklet for them. “I’m sure Maria can do problems 1, 3, and 4. But she will struggle with problems 2 and 5 because she hasn’t mastered those skills yet.” To be effective in supporting learning, teachers need in-depth assessments targeted to the gray areas where they don’t know what their students are thinking.

The following distinctions highlight the differences between large-scale and classroom-level assessments, which imply that notably different assessment strategies are needed.

- Standardized vs. dynamic
- Uniform date vs. ongoing dates
- Independent performance vs. assisted performance
- Delayed feedback vs. immediate feedback
- Stringent requirements for technical accuracy vs. less stringent requirements

For example, it is appropriate to provide hints or to alter the task while assessing for classroom purposes because in so doing the teacher learns what a student can do independently and pinpoints precisely where understanding breaks down.

In contrast to these distinctions, the single most important shared characteristic of large-scale and classroom assessments should be their *alignment* with curriculum standards. Here I do not mean the limited alignment obtained when test publishers show that all of their multiple-choice items can be matched to the categories of a state’s content standards. Rather, I am speaking of the more complete and substantive alignment that occurs when the tasks, problems, and projects in which students are engaged represent the range and depth of what we say we want students to understand and be able to do. Perhaps a better word would be *embodiment*. Assessments at either level should embody and fully represent important learning goals. In science, we can use the National Science Education Standards (NRC 1996) as our learning targets. Assessments at both the large-scale and classroom levels, then, must embody the fundamental concepts, principles, and inquiry skills needed to conduct investigations and evaluate scientific findings as identified by the standards.

For large-scale and classroom assessments to be symbiotic, they must share this common understanding of what it means to do good work in a discipline and ideally

CHAPTER 9

hold a common view of how that expertise develops over time (Pellegrino, Chudowsky, and Glaser 2001). When the conception of curriculum represented by a state's large-scale assessment is at odds with content standards and curricular goals, then the ill effects of teaching to the external, high-stakes test, especially curriculum distortion and nongeneralizable test score gains, will be exaggerated, and it will be more difficult for teachers to use classroom assessment strategies that support conceptual understanding while at the same time working to improve student performance on the state test.

A Model of Classroom Assessment in Support of Learning

Classroom assessment is both formal—involving quizzes, exams, laboratory assignments, and projects—and informal—involving journal entries, observations, and oral questioning. It also serves both formative and summative purposes, depending on whether assessment insights are used to help students take the next steps in learning or to report on the level of achievement attained to date. A rich research literature shows us the dramatic achievement gains that can occur when formative assessment is used (Black and Wiliam 1998; Shepard 2000). Most surveys of practice, however, find that assessment is more often used for grading than for learning.

An ideal model of classroom assessment must address both content and process considerations. The activities in which students engage and the work we ask them to produce determine the real targets of learning regardless of what goals might be stated in curriculum guidelines or lesson plans. Therefore, it is essential that the content of instructional activities capture the big ideas and inquiry skills of the National Science Education Standards. Formative assessments are then embedded within these instructional activities. A student's ability to communicate scientific information might be assessed, for example, when presenting a group's findings to the rest of the class. As implementation of standards-based reform progresses, a bigger challenge is to ensure that summative classroom measures also mirror the standards. Too often, classroom tests measure what is easiest to measure—vocabulary definitions and restatement of laws and principles—rather than, say, the ability to use principles and laws to make a prediction or explain a result. As suggested in the science standards document, improving the content of assessment means “assessing what is most highly valued, assessing rich, well-structured knowledge, and assessing scientific understanding and reasoning” (NRC 1996, 100). Effective classroom assessment also departs from traditional practice in the way assessment is used, becoming much more interactive and a part of the learning process. As documented in research studies (Black and Wiliam 1998; Pellegrino, Chudowsky, and Glaser 2001; Shepard 2000), effective assessment

- activates and builds on prior knowledge,
- makes students' thinking visible and explicit,

- 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. Grade-Eight Performance Task Illustrating Hands-On Instruction and Assessment Focused on “Big Ideas”

Grade Eight Performance Task
Student Instructions for *The Fault Line*

Standing near the railroad track waiting for help to arrive, Joe looked around at the mountains, their rocks all twisted and folded. He'd been on this track for ten years and had never noticed the rocks before! How do they get like that, he wondered.

Directions:
In this investigation you will examine the process that causes rock layers to fold and twist.

Part A

- Follow the directions to set up your plate model.

Set up your plate model as shown in this picture.

1. Check that the lines marked 1993 are lined up.
2. Place about 100 ml of sand on your plate models as in the picture above.
3. Smooth the sand into a thin layer.
4. Slowly slide the Pacific Plate along the 1993 line until “Stop” is even with the edge of the North American Plate.

Slowly slide the Pacific Plate along the 1993 line until “Stop” is even with the edge of the North American Plate.

- Do not move your plate model. Go on to answer the questions on the next page.

Grade Eight Performance Task
Questions for *The Fault Line* –
Components 1 and 2

COMPONENT 1

- After you have worked with the plate model, answer the following questions completely.

1. Describe what happened to the sand when you slid one plate beneath the other.

COMPONENT 2

2. Would the direction of the plate movement affect the formation of the mountains? Explain how.

(Continues on next page.)

Source: Reprinted, by permission, from *A sampler of science assessment*, copyright 1994, California Department of Education, P.O. Box 271, Sacramento, CA 95812-0271.

CHAPTER 9


(Figure 1. continued)

**Grade Eight Performance Task
Student Instructions for *The Fault Line***

Part B


■ Follow the directions below.

Your plate model should look like this.



1. Without disturbing the sand from Part A, line up the lines labeled "1993" on both plates. Your model should look like the drawing above.
2. Slowly slide the paper marked "Pacific Plate" until the lines marked "3093" are lined up.

Slowly slide the paper marked "Pacific Plate" until the lines marked "3093" are lined up.



**Grade Eight Performance Task
Questions for *The Fault Line* –
Components 3 and 4**

COMPONENT 3

■ After you have worked with your plate model in Part B, answer the following questions completely.

3. Describe what happened to the mountains at the plate model boundaries.

COMPONENT 4

4. Look at the map below. The land formations in the **Peninsular National Monument** and the land formations at **Tajima Pass** were once located next to each other. Now they are separated by over 100 miles.



Explain how the **Peninsular National Monument** and **Tajima Pass** were separated from each other.

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 National Science Foundation's Local Systemic Change Initiative) concurrent with development of the state's on-demand test. Here's what they said about the role of assessment in leading instructional change.

What was not appreciated early on is that assessment would become the driver for realizing what it meant to "meet the standards." Initially assessment was seen more as an appendage to curriculum. That was due, in part, to the early recall nature of assessments that contributed minimally in diagnosing student learning, whereas curriculum laid out a road map to follow. Later (after the assessments changed dramatically), it was clearer that assessment indicated whether you reached your destination or not. In other words, the task of the leadership and its team was building a consensus around quality student work in science. This consensus had to be founded upon a different model of student learning than the model most teachers possessed. (Wood and Schmidt 2002)

Program "Diagnosis"

It is popular these days to talk about making large-scale assessments more diagnostic. Colorado's Governor Bill Owens has said that he wants "to turn the annual CSAP exam from just a snapshot of student performance into a diagnostic tool to bring up a child's math, reading, and writing scores" (Whaley 2002). And in the No Child Left Behind Act of 2001, state plans are required to include assessments that "produce individual student interpretive, descriptive, and diagnostic reports, ... that allow parents, teachers, and principals to understand and address the specific academic needs of students." In the next subsection, on individual student "screening," I explain what kinds of information a once-per-year test could reasonably provide on individual students' learning. We should be clear that large-scale assessments cannot be diagnostic of *individual* learning needs in the same way that classroom assessments can be. What large-scale assessments can do is "diagnose" *program* strengths and weaknesses. Typically we refer to this as the program evaluation purpose of large-scale assessment.

When content frameworks used in test construction have sufficient numbers of items by content and processes strands, then it is possible to report assessment results by meaningful subscores. For example, it would be possible for a school to

CHAPTER 9

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. A Released Item from the Delaware State Testing Program (DSTP) with Scoring Tool and Instructional Analysis

LIFE SCIENCE

In the Life Science section of the DSTP [Delaware State Testing Program] students are required to figure out the possible gene pairs that come from two parents. Often this type of genetics word problem will require students to explain how dominant and recessive genes affect the way traits are inherited. One of the released items from spring 2000 DSTP illustrates a genetics question students are asked and what is required to earn complete credit.

Analysis:

After analyzing DSTP results from across the State, it appears that many students are struggling with some of the same genetic concepts. For instance, when expected to construct Punnett squares, students fail to separate the gene pair (alleles) of the parents. This error tends to indicate that students are confused as to how meiosis affects the distribution of chromosomes and subsequently genes. Once the students make this kind of mistake it is impossible for them to determine all the gene pairs for a given characteristic that could come from a set of parents. Furthermore, when students end up with gene combinations (inside the squares) that contain more genetic information than the parents it does not seem to cue them into the fact that they have done something wrong in setting up the Punnett square.

Students also experience difficulty with genetic problems when they are given phenotypic patterns of inheritance and asked to derive information about the genotype of an organism (as in the case of the released problem). Again, if students attempt to construct a Punnett square to answer the question they must first be able to determine the genotype for each of the parent organisms and then separate the alleles across the top and down the side of the square. After completing the simple monohybrid crosses they should then be able to apply their understanding of genetics to explain the relationships between the phenotypes and genotypes of the parents and offspring.

Released Item:

In cats, the gene for short hair (A) is dominant over the gene for long hair (a). A short-haired cat is mated to a long-haired cat, and four kittens are produced, two short-haired and two long-haired. Explain how the two parents could produce these offspring.

Scoring Tool:

Response must indicate in words and/or in a correctly constructed Punnett square the appropriate genotypes of both parents and the predicted offspring. For example:

2 points:

One parent must be heterozygous and therefore, has a 50% chance of giving the short-haired gene and a 50% chance of giving the long-haired gene. The other parent can only give the long-haired gene. Therefore, 50% of the offspring will be long-haired and 50% short-haired. Note: The words "heterozygous" and "homozygous" are not required to receive full credit.

OR

	a	a
A	Aa	Aa
a	aa	aa

(Continues on next page.)

CHAPTER 9

(Figure 2. continued)

OR

Parents: aa x Aa Offspring: 50% Aa 50% aa

1 point: Partially correct response, but some flaws may be included. For example, the student may explain the parent with the dominant gene is carrying the recessive allele, but the combinations inside the Punnett square do not reflect separation of the alleles.

0 points: Incorrect, inappropriate, or incomplete response.

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 knowledge use. Because no one instrument can be a perfectly valid indicator of student achievement, the professional *Standards for Educational and Psychological Testing* (AERA, APA, NCME 1999) require that high-stakes decisions “not be made on the basis of a single test score” (146). While once-per-year state assessments can be made sufficiently accurate to report to parents about a student's level of achievement, they should not be used solely to determine grade-to-grade promotion or high school graduation.

Can these state proficiency tests also be diagnostic at the level of individual students? The answer is no, at least not in the same way that classroom assessments can be diagnostic. Once-per-year survey tests are perhaps better thought of as “screening” instruments, not unlike the health screenings provided in shopping malls. If one indicator shows a danger signal, the first thing you should do is see your doctor for a more complete and accurate assessment.

The same subtest information that is available for program level profiling may also be useful at the level of individual student profiles. Notice, however, that the instructional insights provided earlier by Wood and Schmidt (2002) were based on state patterns for large numbers of students. For individual students, it would be inappropriate to interpret the results of single items, and even subtest peaks and valleys are often not reliably different. Unfortunately, the most commonly reported profiles do not reveal a particular area of weakness, where a student needs more work. Instead, test results most frequently come back with predictable findings of “low on everything” or “high on everything.”

I have explained previously that large-scale tests are too broad to provide (in one or two hours of testing) much detail on a student's knowledge of specific content or skills—such as control of variables, formulating explanations, energy transfer, the effect of heat on chemical reactions, the structure and function of cells, the relationship of diversity and evolution, and so forth. An additional source of difficulty is the match or mismatch between the level of a large-scale test and an individual student's level of functioning. Some state assessment programs are based on basic-skills tests with relatively low-level proficiency standards. Low-level, basic-skills tests provide very little information about the knowledge or knowledge gains of high-performing students. In contrast, in states that built their tests in keeping with the rhetoric of world-class standards, there will be few test items designed to measure the knowledge or knowledge gains of below-grade-level students. NAEP, for example, was designed to measure relatively challenging grade-level content, and therefore yields unreliable total score estimates for students whose performance is below grade level.

I should also emphasize that the item sampling strategies currently used for fill-out test frameworks are not designed with an understanding of learning progressions. The authors of *Knowing What Students Know* (Pellegrino, Chudowsky, and Glaser 2001) explained that current curriculum standards “emphasize *what* students should learn, [but] they do not describe *how* students learn in ways that are maximally useful for guiding instruction and assessment” (256). Thus the fourth-grade NAEP mathematics test is a sample of where students are expected to have gotten by fourth grade, not how they got here. Models of student progression in learning have been developed in research settings, but they have not yet been built into large-scale testing programs. It would be a mistake, therefore, to try to make diagnostic decisions from a fine-grained analysis of test results. Especially, one should not assume that students should be instructed on the easy items on the test before proceeding to the difficult items. Such reasoning would tend to reinforce instructional patterns whereby slower students are assigned rote tasks and more able students are assigned reasoning tasks. A more appropriate instructional strategy, based on comprehension research for example, would ask lower-performing students to reason with simpler material rather than delaying reasoning tasks. The appropriate learning continua needed to plan instructional interventions cannot be inferred by rank ordering the item statistics in a traditional test.

Given the inherent limitations of once-per-year, large-scale assessments, there are only a few ways that large-scale assessments could be made more diagnostic for individual students. Out-of-level testing is one possibility. This strategy would still involve a standard test administration, but students would take a test more appropriate to their performance level (such tests are statistically linked across students to provide an accurate total score for a school even though students are taking different tests). The state of Wyoming is one of a few states experimenting with a more ambitious effort to make state assessments more instructionally relevant. Director of Assessment Scott Marion provided the example in Figure 3 of a curriculum-embedded

CHAPTER 9

assessment. The state, along with the Wyoming Body of Evidence Activities Consortium, developed 15–18 of these assessments in each of four core areas to be used to determine if students have met the state’s graduation standards. Districts are free to use these assessments or to develop their own as long as they meet alignment criteria. The desirable feature of these assessments is that teachers can embed them where they fit best in the high school curriculum, so long as students have had a fair opportunity to learn the necessary material and to demonstrate that learning. “Carmaliticus” could be taken by ninth- or eleventh-grade biology students. Because these tasks exemplify the science standards and are administered in the context of instruction, teachers receive much more immediate and targeted information about student performance than they do from more comprehensive large-scale assessments.

The Wyoming example also illustrates one of the inevitable trade-offs if state assessments were to be made more diagnostic of individual student’s learning. More diagnosis means more testing—so as to gather sufficient data in each skill and content area. More testing can perhaps be justified when it is closely tied to specific units of study. But one could not defend the notion of 5–15 hours of testing for a state-level science assessment. A reasonable principle to govern the design of external tests would be the following: either large-scale assessments should be minimally intrusive because they are being administered for program-level data, or large-scale assessments must be able to demonstrate direct benefit to student learning for additional time spent. For policy makers who want more individual pupil diagnosis, this principle leads to the idea of curriculum-embedded assessments administered at variable times so that results can be used in the context of instruction. The only other alternative is for states to develop curriculum materials with sample assessment tasks

Figure 3. A Curriculum-Embedded Assessment



Science Assessment Activity #7:
Carmaliticus

Introduction: To describe evolutionary change and classification systems, scientists use phylogenetic trees. Pictured [at left] is an example of the organization of a phylogenetic tree into branches.

Science Assessed:

- Knowledge of classification systems and evolutionary change
- Ability to organize organisms into a phylogenetic tree according to observable characteristics

In this activity, you will take on the role of a scientist developing a phylogenetic tree to represent the evolutionary changes and classification of an imaginary organism called a Carmaliticus.

Attached are the 66 **imaginary** organisms, called Carmaliticus. They are organized according to Eras, indicated in the table below. The organisms and the Eras are **not** related to Earth’s geologic time periods or the conditions within earth’s time periods.

(Continues on next page.)

Source: Property of the Wyoming Body of Evidence Activities Consortium and the Wyoming Department of Education. Reprinted with permission.

Eras	Organism #	Time in Millions of Years Ago
Era A	66	245–209
Era B	64–65	208–145
Era C	60–64	144–67
Era D	53–59	66–58
Era E	43–52	57–37
Era F	29–42	36–24
Era G	15–28	23–6
Era H	8–14	5–2
Era I	4–7	1–.1
Recent– Still Living	1–3	Present

Part I – Phylogenetic Tree: Organize the Carmalitici into a phylogenetic tree according to Eras and characteristics of the Carmaliticus. On the tree, link each organism to only one organism from the previous Era, with a line; and indicate the extinction of a branch, with a labeled line.

Part II –Written Explanation: Provide a written report with your phylogenetic tree that includes the following:

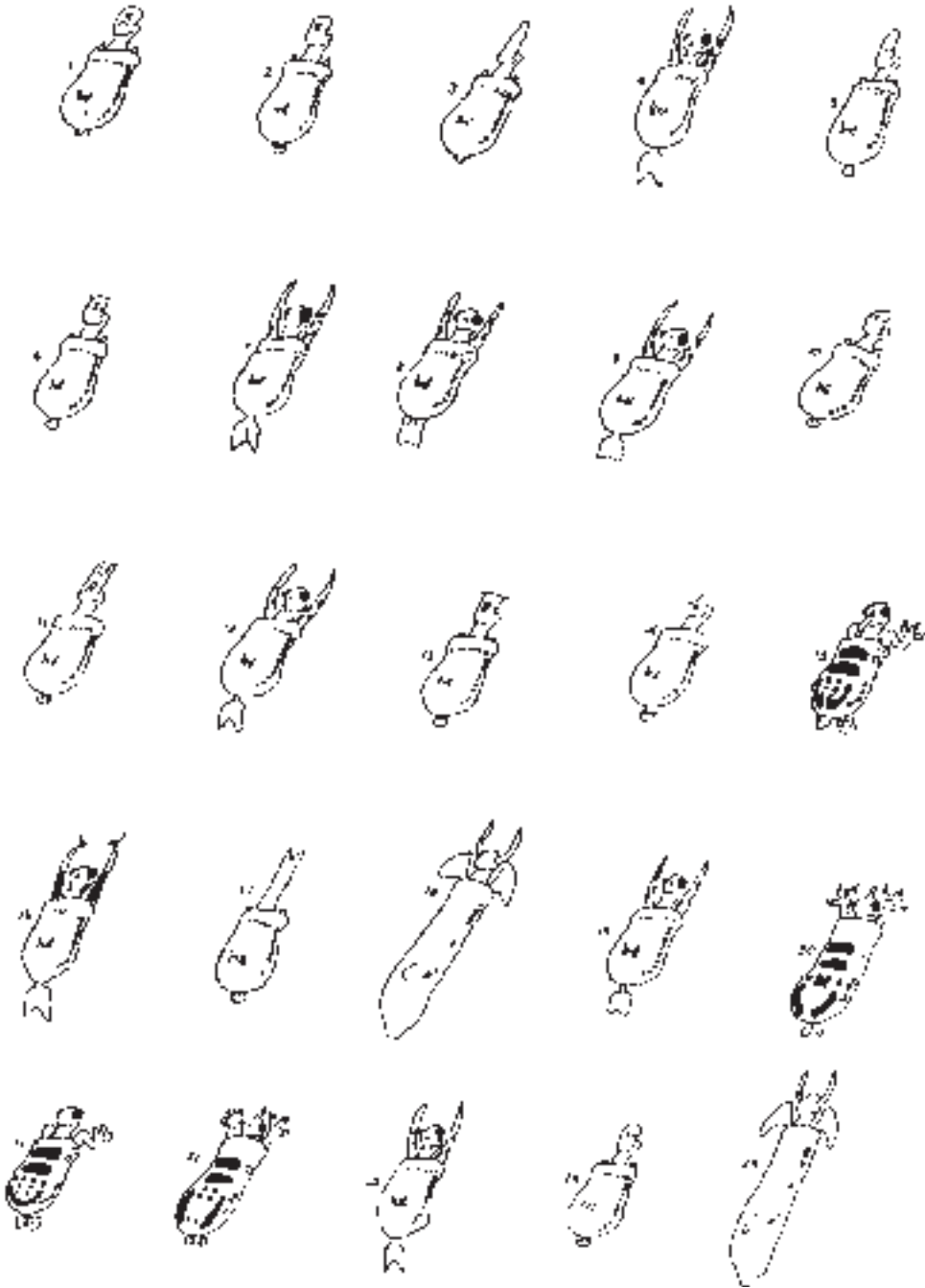
- 1) The reasoning you used to make decisions regarding placement of the Carmaliticus and their branches;
- 2) For *two* branches with *seven or more* Carmaliticus, describe how one organism evolved to another—based on identifiable characteristics of the organisms;
- 3) Possible environments of *four* Eras, supported with characteristics of the organisms that would justify your decisions;
- 4) A comparison of your phylogenetic tree to one other tree produced by a classmate. In your comparison, you are to identify at least two significant differences between your tree and the other tree, including a description about the difference in the organization and characteristics of all of the organisms within at least one branch and a comparison of the branches.

NOTE - Important considerations as you develop your phylogenetic tree:

- a) Consider the organization of the entire tree before attaching the Carmaliticus.
- b) Neatness and spacing will make a difference when you have to examine and explain the individual characteristics and the overall trends of the tree.
- c) Based upon assumptions you make in the development of your tree, it is unlikely that you and another classmate will have an identical tree.
- d) Each organism should only be tied to one other organism from the previous Era.

(Continues on next page.)

CHAPTER 9



(Continues on next page.)



(Continues on next page.)

CHAPTER 9



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

CHAPTER 9

wrong or right for different reasons. Making explicit an array of student thinking around a question forces the teachers to think about the implications for instructional practice. This characterization of student learning resonated more with teachers at a gut level than the daunting but somewhat intriguing collection of student thinking documented in the research base. (Wood and Schmidt 2002)

Dr. Maryellen Harmon, a consultant to Wood and Schmidt's (2002) project, required that teachers who were developing summative assessment tasks take time to write out what each item is measuring and also to write out their own "elegant answers" to each question before developing the scoring criteria and rubric. From these "academic exercises," teachers found they could catch flaws in their own thinking and sometimes reconsidered whether the target was even worth measuring in the first place. They also became aware of how students would struggle when they themselves could not agree on what was being asked or required for a complete response. During pilot testing, Dr. Harmon also coached teachers to learn from student responses and not always blame the students when they couldn't respond. Although Wood and Schmidt focused on whether learning from the summative assessment project could be generalized to developing items for the state test, these skills could as likely be generalized to developing better classroom assessment. As a result of these experiences, "teachers were much more willing to pilot potential DSTP [Delaware Student Testing Program] test items prior to submitting them and were more aware of how to interpret student work. Many of these teachers now write out a "what this test measures ..." when they construct an item for the state test. They are much less likely to blame a student for an unanticipated response and more likely to reexamine their question and rubric."

The assessment development process and pilot testing experiences described by Wood and Schmidt (2002) show us the power of real professional development opportunities as compared to merely receiving student scores from a state test. "When lead teachers had to score student work from a unit that was just taught, teachers had to evaluate both the extent to which students had acquired certain concepts as well as reflect on their own teaching strategies for particular lessons." For example, "teachers had assumed that students could trace the path of electricity in a complete circuit. When their own data contradicted their assumptions, they realized the need to address this learning in another way with their students." Most tellingly, teachers had to face the dissonance between what they had taught and what students had learned:

After all, these teachers knew that good science was happening in their classrooms—they were using NSF materials, had undergone the training on the modules, and were comfortable with the content knowledge now and employing inquiry-based strategies. The students were active in their learning and enjoyed the lessons immensely. Imagine the impact of data that confronts and challenges

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 assessments built in. And, as illustrated by Wood and Schmidt’s (2002) experiences, they need extended support while attempting to use these materials and draw inferences about how to improve instruction.

Conclusion: Impediments and Recommendations

The single most important requirement to increase the likelihood that large-scale assessments will contribute positively to student learning is to improve the substance of what is assessed. If large-scale assessments were to embody important learning goals—not only inquiry skills, but also the important big ideas in content areas, geological time scale, photosynthesis, why electric current is different from “flowing” water, why we isolate smallpox patients and not AIDS patients—then other aspects of the assessment, such as program evaluation profiles, released item insights, and professional development, can also be used to improve instruction. In *Knowing What Students Know*, Pellegrino, Chudowsky, and Glaser (2001) argued that for an assessment system to support learning, it has to have the feature of *coherence*. That means that classroom and external assessments have to share the same or compatible underlying models of student learning; otherwise, as in the present-day system, they will work at cross purposes.

While a large-scale assessment might be based on a model of learning that is coarser than that underlying the assessments used in classrooms, the conceptual base for the large-scale assessment should be a broader version of one that makes sense at the finer-grained level (Mislevy 1996). In this way, the external assessment results will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment. (Pellegrino, Chudowsky, and Glaser 2001, 255–56)

CHAPTER 9

In attempting to pursue this vision of an ideal assessment system, science educators should be aware of several potential obstacles:

- Cost
- The No Child Left Behind Act's mandate for testing every pupil (with no out-of-level testing)
- Technical standards and legal protections
- Curriculum control
- Lack of trust of teachers as evaluators
- Beliefs held by policy makers about standards-based reform
- Mechanical data systems

Substantively ambitious assessments can be developed and scored reliably for large-scale purposes, but they invariably cost more than machine-scored, multiple-choice tests. Passage of the No Child Left Behind Act has so markedly increased the amount of testing required that we are likely to see a continuing decline in the substantive quality of large-scale tests, because state agencies often cannot afford to do better. Science educators have the advantage that science will be assessed less frequently than reading and mathematics, and therefore, it is more feasible to advocate for high-quality science assessments. Technical standards and legal protections also tend to work against the quality of assessments simply because trivial things are more easily measured consistently. Therefore, the case will have to be made as to why better assessments are worth the investment (i.e., why it's worth it to spend the extra money to measure important things consistently).

Other obstacles to assessment reform include issues of curriculum control and lack of trust of teachers as evaluators. Successful implementations of substantively ambitious assessments, such as the New Standards Project (1997) and the Educational Testing Service's Pacesetter program, have moved much closer to curriculum development than traditional test construction, which merely collected test items. A problem arises then, when states make the tests and districts control curriculum, about how to achieve the kind of coherence envisioned by Pellegrino, Chudowsky, and Glaser (2001). Similarly, teachers gain more from assessments when they are involved in providing data, and teacher participation makes it more likely that assessments can include extended tasks grounded in classroom work. Therefore, including portfolio and project data would increase the validity and meaningfulness of a large-scale assessment. But because of distrust, which motivates the accountability movement in the first place, proponents of substantively richer assessments will have to think of safeguards, such as score moderation schemes that verify the accuracy of teacher-reported data, to counter the claim that teachers might misrepresent student achievement.

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?

CHAPTER 9

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 Researcher* 18: 27–32.
- Mislevy, R. J. 1996. Test theory reconceived. *Journal of Educational Measurement* 33(4): 379–416.
- National Research Council (NRC). 1996. *National science education standards*. Washington, DC: National Academy Press.
- New Standards Project. 1997. *Performance standards: English language arts, mathematics, science, applied learning*. Vols. 1–3. Washington, DC: National Center for Education Statistics and the University of Pittsburgh.
- Pellegrino, J. W., N. Chudowsky, and R. Glaser. 2001. *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Resnick, L. B., and D. P. Resnick. 1992. Assessing the thinking curriculum: New tools for educational reform. In *Changing assessments: Alternative views of aptitude, achievement, and instruction*. Boston: Kluwer.
- Sadler, R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18: 119–44.
- Shepard, L. A. 2000. The role of assessment in a learning culture. *Educational Researcher* 29(7): 4–14.
- Smith, M. S., and J. O'Day. 1990. Systemic school reform. In *Politics of education association yearbook 1990*, 233–67. London: Taylor and Francis.
- Stipek, D. J. 1996. Motivation and instruction. In *Handbook of educational psychology*, eds. D. C. Berliner and R. C. Calfee, 85–113. New York: Macmillan.
- Whaley, M. 2002. Owens looks to broaden CSAP focus: Governor wants student-performance test to become tool for individual improvement. *Denver Post*, 14 March.
- Wood, R., and J. Schmidt. 2002. History of the development of Delaware Comprehensive Assessment Program in Science. Unpublished memorandum.