
observe
Learn
Interact
Change
Student achievement can be interpreted only in light of the quality of the programs they have experienced.

Assessment in Science Education

The assessment standards provide criteria to judge progress toward the science education vision of scientific literacy for all. The standards describe the quality of assessment practices used by teachers and state and federal agencies to measure student achievement and the opportunity provided students to learn science. By identifying essential characteristics of exemplary assessment practices, the standards serve as guides for developing assessment tasks, practices, and policies. These standards can be applied equally to the assessment of students, teachers, and programs; to summative and formative assessment practices; and to classroom assessments as well as large-scale, external assessments. This chapter begins with an introduction that describes the components of the assessment process and a contemporary view of measurement theory and practice. This introduction

is followed by the assessment standards and then by discussions of some ways teachers use assessments and some characteristics of assessments conducted at the district, state, and national levels. The chapter closes with two sample assessment tasks, one to probe students' understanding of the natural world and another to probe their ability to inquire.

The assessment process is an effective tool for communicating the expectations of the science education system to all concerned with science education.

In the vision described by the *National Science Education Standards*, assessment is a primary feedback mechanism in the science education system. For example, assessment data provide students with feedback on how well they are meeting the expectations of their teachers and parents, teachers with feedback on how well their students are learning, districts with feedback on the effectiveness of their teachers and programs, and policy makers with feedback on how well policies are working. Feedback leads to changes in the science education system by stimulating changes in policy, guiding teacher professional development, and encouraging students to improve their understanding of science.

The assessment process is an effective tool for communicating the expectations of the science education system to all concerned with science education. Assessment practices and policies provide operational definitions of what is important. For example, the use of an extended inquiry for an assessment task signals what students are to learn, how teachers are to teach, and where resources are to be allocated.

Assessment is a systematic, multistep process involving the collection and interpretation of educational data. The four components of the assessment process are detailed in Figure 5.1.

As science educators are changing the way they think about good science education, educational measurement specialists are acknowledging change as well. Recognition of the importance of assessment to contemporary educational reform has catalyzed research, development, and implementation of new methods of data collection along with new ways of judging data quality. These changes in measurement theory and practice are reflected in the assessment standards.

In this new view, assessment and learning are two sides of the same coin. The methods used to collect educational data define in measurable terms what teachers should teach and what students should learn. And when students engage in an assessment exercise, they should learn from it.

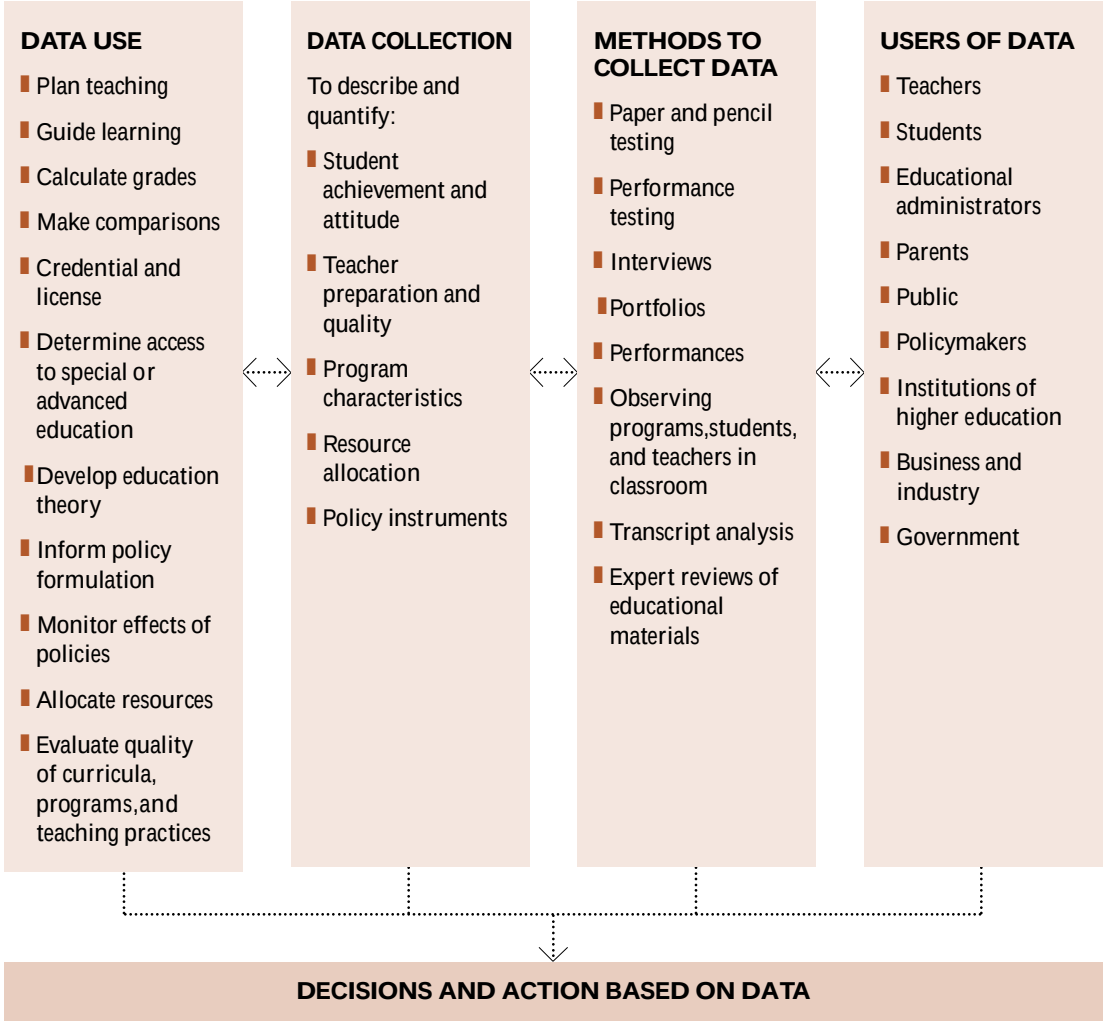
This view of assessment places greater confidence in the results of assessment procedures that sample an assortment of variables using diverse data-collection methods, rather than the more traditional sampling of one variable by a single method. Thus, all aspects of science achievement—ability to inquire, scientific understanding of the natural world, understanding of the nature and utility of science—are measured using multiple methods such as performances and portfolios, as well as conventional paper-and-pencil tests.

The assessment standards include increased emphasis on the measurement of

See Assessment Standard B

FIGURE 5.1. COMPONENTS OF THE ASSESSMENT PROCESS

The four components can be combined in numerous ways. For example, teachers use student achievement data to plan and modify teaching practices,and business leaders use per capita educational expenditures to locate businesses. The variety of uses,users,methods,and data contributes to the complexity and importance of the assessment process.



opportunity to learn. Student achievement can be interpreted only in light of the quality of the programs they have experienced.

Another important shift is toward "authentic assessment." This movement calls for exercises that closely approximate the intended outcomes of science education. Authentic assessment exercises require students to apply scientific knowledge and reasoning to situations similar to those they will encounter in the world outside the classroom, as well as to situations that approximate how scientists do their work.

Another conceptual shift within the educational measurement area that has significant implications for science assessment involves validity. Validity must be concerned not only with the technical quality of educational data, but also with the social and educational consequences of data interpretation.

An important assumption underlying the assessment standards is that states and local districts can develop mechanisms to measure students' achievement as specified in the content standards and to measure the opportunities for learning science as specified in the program and system standards. If the principles in the assessment standards are followed, the information resulting from new modes of assessment applied locally can have common meaning and value in terms of the national standards, despite the use of different assessment procedures and instruments in different locales. This contrasts with the traditional view of educational measurement that allows for comparisons only when they are based on parallel forms of the same test.

The Standards

ASSESSMENT STANDARD A:
Assessments must be consistent with the decisions they are designed to inform.

- **Assessments are deliberately designed.**
- **Assessments have explicitly stated purposes.**
- **The relationship between the decisions and the data is clear.**
- **Assessment procedures are internally consistent.**

The essential characteristic of well-designed assessments is that the processes used to collect and interpret data are consistent with the purpose of the assessment. That match of purpose and process is achieved through thoughtful planning that is available for public review.

ASSESSMENTS ARE DELIBERATELY DESIGNED. Educational data profoundly influence the lives of students, as well as the people and institutions responsible for science education. People who must use the results of assessments to make decisions and take actions, as well as those who are affected by the decisions and actions, deserve assurance that assessments are carefully conceptualized. Evidence of careful conceptualization is found in written plans for assessments that contain

- Statements about the purposes that the assessment will serve.
- Descriptions of the substance and technical quality of the data to be collected.
- Specifications of the number of students or schools from which data will be obtained.

- Descriptions of the data-collection method.
- Descriptions of the method of data interpretation.
- Descriptions of the decisions to be made, including who will make the decisions and by what procedures.

ASSESSMENTS HAVE EXPLICITLY STATED PURPOSES. Conducting assessments is a resource-intensive activity. Routine assessments in the classroom place considerable demands on the time and intellectual resources of teachers and students. Large-scale assessments, such as those conducted by districts, states, and the federal government, require tremendous human and fiscal expenditures. Such resources should be expended only with the assurance that the decisions and actions that follow will increase the scientific literacy of the students—an assurance that can be made only if the purpose of the assessment is clear.

THE RELATIONSHIP BETWEEN THE DECISIONS AND THE DATA IS CLEAR. Assessments test assumptions about relationships among educational variables. For example, if the purpose is to decide if a school district's management system should be continued, assessment data might be collected about student achievement. This choice of assessment would be based on the following assumed relationship: the management system gives teachers responsibility for selecting the science programs, teachers have an incentive to implement effectively the programs they select, and effective implementation improves science achievement. The relationship between the decision to be made and the data to be collected is specified.

See Teaching Standard F

ASSESSMENT PROCEDURES NEED TO BE INTERNALLY CONSISTENT. For an assessment to be internally consistent, each component must be consistent with all others. A link of inferences must be established and reasonable alternative explanations eliminated. For example, in the district management example above, the relationship between the management system and student achievement is not adequately tested if student achievement is the only variable measured. The extent to which the management system increased teacher responsibility and led to changes in the science programs that could influence science achievement must also be measured.

ASSESSMENT STANDARD B: Achievement and opportunity to learn science must be assessed.

- **Achievement data collected focus on the science content that is most important for students to learn.**
- **Opportunity-to-learn data collected focus on the most powerful indicators.**
- **Equal attention must be given to the assessment of opportunity to learn and to the assessment of student achievement.**

ACHIEVEMENT DATA COLLECTED FOCUS ON THE SCIENCE CONTENT THAT IS MOST IMPORTANT FOR STUDENTS TO LEARN. The content standards define the science all students will come to understand. They portray the outcomes of science education as rich and varied, encompassing

- The ability to inquire.
- Knowing and understanding scientific facts, concepts, principles, laws, and theories.

The Insect and the Spider

Titles in this example emphasize some of the components of the assessment process. In the vision of science education described in the Standards, *teaching often cannot be distinguished from assessment. In this example, Ms. M. uses information from observations of student work and discussion to change classroom practice to improve student understanding of complex ideas. She has a repertoire of analogies, questions, and examples that she has developed and uses when needed. The students develop answers to questions about an analogy using written and diagrammatic representations. The administrator recognizes that teachers make plans but adapt them and provided Ms. M. with an opportunity to explain the reasoning supporting her decision.*

[This example highlights some elements of Teaching Standard A and B; Assessment Standard A, 5-8 Content Standard B, and Program Standard F.]

SCIENCE CONTENT: The 5-8 Physical Science Content Standard includes an understanding of motions and forces. One of the supporting ideas is that the motion of an object can be described by the change in its position with time.

ASSESSMENT ACTIVITY: Students respond to questions about frames of reference with extended written responses and diagrams.

ASSESSMENT TYPE: This is an individual extended response exercise embedded in teaching.

ASSESSMENT PURPOSE: The teacher uses the information from this activity to improve the lesson.

DATA: Students' written responses.
Teacher's observations.

CONTEXT: A seventh-grade class is studying the motion of objects. One student, describing his idea about motion and forces, points to a book on the desk and says "right now the book is not moving." A second student interrupts, "Oh, yes it is. The book is on the desk, the desk is on the floor, the floor is a part of the building, the building is sitting on the Earth, the Earth is rotating on its axis and revolving around the Sun, and the whole solar system is moving through the Milky Way." The second student sits back with a self-satisfied smile on her face. All discussion ceases.

Ms. M. signals time and poses the following questions to the class. Imagine an insect and a spider on a lily pad floating down a stream. The spider is walking around the edge of lily pad. The insect is sitting in the middle of the pad watching the spider. How would the insect describe its own motion? How would the insect describe the spider's motion? How would a bird sitting on the edge of the stream describe the motion of the insect and the spider? After setting the class to work discussing the questions, the teacher walks around the room listening to the discussions. Ms. M. asks the students to write answers to the questions she posed; she suggests that the students use diagrams as a part of the responses.

The school principal had been observing Ms. M. during this class and asked her to explain why she had not followed her original lesson plan. Ms. M. explained that the girl had made a similar statement to the class twice before. Ms. M. realized that the girl was not being disruptive but was making a legitimate point that the other members of the class were not grasping. So Ms. M. decided that continuing with the discussion of motions and forces would not be fruitful until the class had developed a better concept of frame of reference. Her questions were designed to help the students realize that motion is described in terms of some point of reference. The insect in the middle of lily pad would describe its motion and the motion of the spider in terms of its reference frame, the lily pad. In contrast, the bird watching from the edge of the stream would describe the motion of the lily pad and its passengers in terms of its reference frame, namely the ground on which it was standing. Someone on the ground observing the bird would say that the bird was not in motion, but an observer on the moon would have a different answer.

- The ability to reason scientifically.
- The ability to use science to make personal decisions and to take positions on societal issues.
- The ability to communicate effectively about science.

This assessment standard highlights the complexity of the content standards while addressing the importance of collecting data on all aspects of student science achievement. Educational measurement theory and practice have been well developed primarily to measure student knowledge about subject matter; therefore, many educators and policy analysts have more confidence in instruments designed to measure a student's command of information about science than in instruments designed to measure students' understanding of the natural world or their ability to inquire. Many current science achievement tests measure "inert" knowledge—discrete, isolated bits of knowledge—rather than "active" knowledge—knowledge that is rich and well-structured. Assessment processes that include all outcomes for student achievement must probe the extent and organization of a student's knowledge. Rather than checking whether students have memorized certain items of information, assessments need to probe for students' understanding, reasoning, and the utilization of knowledge. Assessment and learning are so closely related that if all the outcomes are not assessed, teachers and students likely will redefine their expectations for learning science only to the outcomes that are assessed.

See Content Standards B, C, and D (all grade levels)

See the principal *Learning science is an active process* in Chapter 2

OPPORTUNITY-TO-LEARN DATA COLLECTED FOCUS ON THE MOST POWERFUL INDICATORS. The system, program, teaching, and professional development standards portray the conditions that must exist throughout the science education system if all students are to have the opportunity to learn science.

At the classroom level, some of the most powerful indicators of opportunity to learn are teachers' professional knowledge, including content knowledge, pedagogical knowledge, and understanding of students; the extent to which content, teaching, professional development, and assessment are coordinated; the time available for teachers to teach and students to learn science; the availability of resources for student inquiry; and the quality of educational materials available. The teaching and program standards define in greater detail these and other indicators of opportunity to learn.

Some indicators of opportunity to learn have their origins at the federal, state, and district levels and are discussed in greater detail in the systems standards. Other powerful indicators of opportunity to learn beyond the classroom include per-capita educational expenditures, state science requirements for graduation, and federal allocation of funds to states.

Compelling indicators of opportunity to learn are continually being identified, and ways to collect data about them are being designed. Measuring such indicators presents many technical, theoretical, economic, and social challenges, but those challenges do not obviate the responsibility of moving forward on implementing and assessing opportunity to learn. The assessment standards call for a policy-level commitment of the resources necessary for research and development related to assessing opportunity to learn. That commitment includes the

development of the technical skills to assess opportunity to learn among science education professionals, including teachers, supervisors, administrators,and curriculum developers.

EQUAL ATTENTION MUST BE GIVEN TO THE ASSESSMENT OF OPPORTUNITY TO LEARN AND TO THE ASSESSMENT OF STUDENT ACHIEVEMENT. Students cannot be held accountable for achievement unless they are given adequate opportunity to learn science. Therefore, achievement and opportunity to learn science must be assessed equally.

See Program Standard E and System Standard E

ASSESSMENT STANDARD C:
The technical quality of the data collected is well matched to the decisions and actions taken on the basis of their interpretation.

- **The feature that is claimed to be measured is actually measured.**
- **Assessment tasks are authentic.**
- **An individual student's performance is similar on two or more tasks that claim to measure the same aspect of student achievement.**
- **Students have adequate opportunity to demonstrate their achievements.**
- **Assessment tasks and methods of presenting them provide data that are sufficiently stable to lead to the same decisions if used at different times.**

Standard C addresses the degree to which the data collected warrant the decisions and actions that will be based on them. The quality of the decisions and the appropriateness of resulting action are limited by the quality of the data. The more serious the consequences for students or teachers, the greater confidence those making the decisions must have in the technical quality of the data. Confidence is gauged by the quality of the assessment process and the consistency of the measurement over alternative assessment processes. Judgments about confidence are based on several different indicators, some of which are discussed below.

THE FEATURE THAT IS CLAIMED TO BE MEASURED IS ACTUALLY MEASURED. The content and form of an assessment task must be congruent with what is supposed to be measured. This is "validity." For instance, if an assessment claims to measure students' ability to frame questions for conducting scientific inquiry and to design an inquiry to address the questions, a short-answer format would not be an appropriate task. Requiring students to pose questions and design inquiries to address them would be an appropriate task. However, if the purpose of an assessment task is to measure students' knowledge of the characteristics that distinguish groups of minerals, a multiple-choice format might be suitable as well as efficient.

The content and form of an assessment task must be congruent with what is supposed to be measured.

ASSESSMENT TASKS ARE AUTHENTIC. When students are engaged in assessment tasks that are similar in form to tasks in which they will engage in their lives outside the classroom or are similar to the activities of scientists, great confidence can be attached to the data collected. Such assessment tasks are authentic.

Classroom assessments can take many forms, including observations of student performance during instructional activities; interviews; formal performance tasks; portfolios; investigative projects; written reports; and multiple choice, short-answer, and essay examinations. The relationship of some of those forms of assessment tasks to the goals of science education are not as obvious as others. For instance, a student's ability to obtain and evaluate scientific information might be measured using a short-answer test to identify the sources of high-quality scientific information about toxic waste. An alternative and more authentic method is to ask the student to locate such information and develop an annotated bibliography and a judgment about the scientific quality of the information.

AN INDIVIDUAL STUDENT'S PERFORMANCE IS SIMILAR ON TWO OR MORE TASKS THAT CLAIM TO MEASURE THE SAME ASPECT OF STUDENT ACHIEVEMENT. This is one aspect of reliability. Suppose that the purpose of an assessment is to measure a student's ability to pose appropriate questions.A student might be asked to pose questions in a situation set in the physical sciences. The student's performance and the task are consistent if the performance is the same when the task is set in the context of the life sciences, assuming the student has had equal opportunities to learn physical and life sciences.

STUDENTS HAVE ADEQUATE OPPORTUNITIES TO DEMONSTRATE THEIR ACHIEVEMENTS. For decision makers to have confidence in assessment data, they need assurance that students have had the opportunity to demonstrate their full understanding and ability. Assessment tasks must be developmentally appropriate, must be set in contexts that are familiar to the students, must not require reading skills or vocabulary that are inappropriate to the students' grade level, and must be as free from bias as possible.

ASSESSMENT TASKS AND THE METHODS OF PRESENTING THEM PROVIDE DATA THAT ARE SUFFICIENTLY STABLE TO LEAD TO THE SAME DECISIONS IF USED AT DIFFERENT TIMES. This is another aspect of reliability, and is especially important for large-scale assessments, where changes in performance of groups is of interest. Only with stable measures can valid inferences about changes in group performance be made.

Although the confidence indicators discussed above focus on student achievement data, an analogous set of confidence indicators can be generated for opportunity to learn. For instance, teacher quality is an indicator of opportunity to learn. Authenticity is obtained if teacher quality is measured by systematic observation of teaching performance by qualified observers. Confidence in the measure is

Assessment tasks must be developmentally appropriate, must be set in contexts that are familiar to the students, must not require reading skills or vocabulary that are inappropriate to the students' grade level, and must be as free from bias as possible.

See Teaching Standard C

achieved when the number of observations is large enough for a teacher to exhibit a full range of teaching knowledge and skill. Consistency of performance is also established through repeated observations.

Data-collection methods can take many forms. Each has advantages and disadvantages. The choice among them is usually constrained by tradeoffs between the type, quality, and amount of information gained, and the time and resources each requires. However, to serve the intended purpose, the choice of assessment form should be consistent with what one wants to measure and to infer. It is critical that the data and their method of collection yield information with confidence levels consistent with the consequences of its use. Public confidence in educational data and their use is related to technical quality. This public confidence is influenced by the extent to which technical quality has been considered by educators and policy makers and the skill with which they communicate with the public about it.

The choice of assessment form should be consistent with what one wants to measure and to infer.

ASSESSMENT STANDARD D:
Assessment practices must be fair.

- **Assessment tasks must be reviewed for the use of stereotypes, for assumptions that reflect the perspectives or experiences of a particular group, for language that might be offensive to a particular group, and for other features that might distract students from the intended task.**
- **Large-scale assessments must use statistical techniques to identify potential bias among subgroups.**
- **Assessment tasks must be appropriately modified to accommodate the needs of students with physical disabilities, learning disabilities, or limited English proficiency.**
- **Assessment tasks must be set in a variety of contexts, be engaging to students with different interests and experiences, and must not assume the perspective or experience of a particular gender, racial,or ethnic group.**

A premise of the *National Science Education Standards* is that all students should have access to quality science education and should be expected to achieve scientific literacy as defined by the content standards. It follows that the processes used to assess student achievement must be fair to all students. This is not only an ethical requirement but also a measurement requirement. If assessment results are more closely related to gender or ethnicity than to the preparation received or the science understanding and ability being assessed, the validity of the assessment process is questionable.

See Program Standard E and System Standard E

ASSESSMENT TASKS MUST BE REVIEWED FOR THE USE OF STEREOTYPES, FOR ASSUMPTIONS THAT REFLECT THE PERSPECTIVES OR EXPERIENCES OF A PARTICULAR GROUP, FOR LANGUAGE THAT MIGHT BE OFFENSIVE TO A PARTICULAR GROUP, AND FOR OTHER FEATURES THAT MIGHT DISTRACT STUDENTS FROM THE INTENDED TASK. Those who plan and implement science assessments must pay deliberate attention to issues of fairness.

The concern for fairness is reflected in the procedures used to develop assessment tasks, in the content and language of the assessment tasks, in the processes by which students are assessed, and in the analyses of assessment results.

LARGE-SCALE ASSESSMENTS MUST USE STATISTICAL TECHNIQUES TO IDENTIFY POTENTIAL BIAS AMONG SUBGROUPS. Statistical techniques require that both sexes and different racial and ethnic backgrounds be included in the development of large-scale assessments. Bias can be determined with some certainty through the combination of statistical evidence and expert judgment. For instance, if an exercise to assess understanding of inertia using a flywheel results in differential performance between females and males, a judgment that the exercise is biased might be plausible based on the assumption that males and females have different experiences with flywheels.

ASSESSMENT TASKS MUST BE MODIFIED APPROPRIATELY TO ACCOMMODATE THE NEEDS OF STUDENTS WITH PHYSICAL DISABILITIES, LEARNING DISABILITIES, OR LIMITED ENGLISH PROFICIENCY. Whether assessments are large scale or teacher conducted, the principle of fairness requires that data-collection methods allow students with physical disabilities, learning disabilities, or limited English proficiency to demonstrate the full extent of their science knowledge and skills.

ASSESSMENT TASKS MUST BE SET IN A VARIETY OF CONTEXTS, BE ENGAGING TO STUDENTS WITH DIFFERENT INTERESTS AND EXPERIENCES, AND MUST NOT ASSUME THE PERSPECTIVE OR EXPERIENCE OF A PARTICULAR GENDER, RACIAL, OR ETHNIC GROUP. The requirement that assessment exercises be authentic and thus in context increases the likelihood that all tasks have some degree of bias for some population of students. Some contexts will have more appeal to males and others to females. If, however, assessments employ a variety of tasks, the collection will be "equally unfair" to all. This is one way in which the deleterious effects of bias can be avoided.

ASSESSMENT STANDARD E:
The inferences made from assessments about student achievement and opportunity to learn must be sound.

- **When making inferences from assessment data about student achievement and opportunity to learn science, explicit reference needs to be made to the assumptions on which the inferences are based.**

Even when assessments are well planned and the quality of the resulting data high, the interpretations of the empirical evidence can result in quite different conclusions. Making inferences involves looking at empirical data through the lenses of theory, personal beliefs, and personal experience. Making objective inferences is extremely difficult, partly because individuals are not always aware of their assumptions. Consequently, confidence in the validity of inferences requires explicit reference to the assumptions on which those inferences are based.

For example, if the science achievement on a large-scale assessment of a sample of students from a certain population is high, several conclusions are possible. Students

from the population might be highly motivated; or because of excellent instruction, students from the population might have greater opportunity to learn science; or the test might be biased in some way in favor of the students. Little confidence can be placed in any of these conclusions without clear statements about the assumptions and a developed line of reasoning from the evidence to the conclusion. The level of confidence in conclusions is raised when those conducting assessments have been well trained in the process of making inferences from educational assessment data. Even then, the general public, as well as professionals, should demand open and understandable descriptions of how the inferences were made.

Assessments Conducted by Classroom Teachers

Teachers are in the best position to put assessment data to powerful use. In the vision of science education described by the *Standards*, teachers use the assessment data in many ways. Some of the ways teachers might use these data are presented in this section.

IMPROVING CLASSROOM PRACTICE

See Teaching Standard C

Teachers collect information about students' understanding almost continuously and make adjustments to their teaching on the basis of their interpretation of that information. They observe critical incidents in the classroom, formulate hypotheses about the causes of those incidents, question students to test their hypotheses, interpret student's responses, and adjust their teaching plans.

PLANNING CURRICULA

Teachers use assessment data to plan curricula. Some data teachers have collected themselves; other data come from external sources. The data are used to select content, activities,and examples that will be incorporated into a course of study, a module, a unit, or a lesson. Teachers use the assessment data to make judgments about

- The developmental appropriateness of the science content.
- Student interest in the content.
- The effectiveness of activities in producing the desired learning outcomes.
- The effectiveness of the selected examples.
- The understanding and abilities students must have to benefit from the selected activities and examples.

Planning for assessment is integral to instruction. Assessments embedded in the curriculum serve at least three purposes: to determine the students' initial understandings and abilities, to monitor student progress,and to collect information to grade student achievement. Assessment tasks used for those purposes reflect what students are expected to learn; elicit the full extent of students' understanding; are set in a variety of contexts; have practical, aesthetic, and heuristic value; and have meaning outside the classroom. Assessment tasks also provide important clues to students about what is important to learn.

DEVELOPING SELF-DIRECTED LEARNERS

Students need the opportunity to evaluate and reflect on their own scientific understanding and ability. Before students can do this, they need to understand the goals for learning science. The ability to self-assess understanding is an essential tool for self-directed learning. Through self-reflection, students clarify ideas of what they are supposed to learn. They begin to internalize the expectation that they can learn science. Developing self-assessment skills is an ongoing process throughout a student's school career, becoming increasingly more sophisticated and self-initiated as a student progresses.

When teachers treat students as serious learners and serve as coaches rather than judges, students come to understand and apply standards of good scientific practice.

Conversations among a teacher and students about assessment tasks and the teacher's evaluation of performance provide students with necessary information to assess their own work. In concert with opportunities to apply it to individual work and to the work of peers, that information contributes to the development of students' self-assessment skills. By developing these skills, students become able to take responsibility for their own learning.

Teachers have communicated their assessment practices, their standards for performance, and criteria for evaluation to students when students are able to

- Select a piece of their own work to provide evidence of understanding of a scientific concept, principle, or law—or their ability to conduct scientific inquiry.
- Explain orally, in writing, or through illustration how a work sample provides evidence of understanding.
- Critique a sample of their own work using the teacher's standards and criteria for quality.
- Critique the work of other students in constructive ways.

Involving students in the assessment process increases the responsibilities of the teacher. Teachers of science are the representatives of the scientific community in their classrooms; they represent a culture and a way of thinking that might be quite unfamiliar to students. As representatives, teachers are expected to model reflection, fostering a learning environment where students review each others' work, offer suggestions, and challenge mistakes in investigative processes, faulty reasoning, or poorly supported conclusions.

A teacher's formal and informal evaluations of student work should exemplify scientific practice in making judgments. The standards for judging the significance, soundness, and creativity of work in professional scientific work are complex, but they are not arbitrary. In the work of classroom learning and investigation, teachers represent the standards of practice of the scientific community. When teachers treat students as serious learners and serve as coaches rather than judges, students come to understand and apply standards of good scientific practice.

REPORTING STUDENT PROGRESS

An essential responsibility of teachers is to report on student progress and achievement

See Teaching Standard C

to the students themselves, to their colleagues, to parents and to policy makers. Progress reports provide information about

- The teacher's performance standards and criteria for evaluation.
- A student's progress from marking period to marking period and from year to year.
- A student's progress in mastering the science curriculum.
- A student's achievement measured against standards-based criteria.

See System Standards A and B

Each of these issues requires a different kind of information and a different mode of assessment.

Especially challenging for teachers is communicating to parents and policy makers the new methods of gathering information that are gaining acceptance in schools. Parents and policy makers need to be reassured that the newer methods are not only as good as, but better than,those used when they were in school. Thus, in developing plans for assessment strategies to compile evidence of student achievement, teachers demonstrate that alternative forms of data collection and methods of interpreting them are as valid and reliable as the familiar short-answer test.

The purported objectivity of short-answer tests is so highly valued that newer modes of assessment such as portfolios, performances,and essays that rely on apparently more subjective scoring methods are less trusted by people who are not professional educators. Overcoming this lack of trust requires that teachers use assessment plans for monitoring student progress and for grading. Clearly relating assessment tasks and products of student work to the valued goals of science education is integral to assessment plans. Equally important is that the plans have explicit criteria for judging the quality of students' work that policy makers and parents can understand.

RESEARCHING TEACHING PRACTICES

See Professional Development Standards B and C

Master teachers engage in practical inquiry of their own teaching to identify conditions that promote student learning and to understand why certain practices are effective. The teacher as a researcher engages in assessment activities that are similar to scientific inquiries when collecting data to answer questions about effective teaching practices. Engaging in classroom research means that teachers develop assessment plans that involve collecting data about students' opportunities to learn as well as their achievement.

Assessments Conducted at the District, State, and National Levels

See System Standards A and B

Science assessments conducted by district, state, and national authorities serve similar purposes and are distinguished primarily by scale—that is, by the number of students, teachers, or schools on which data are collected.

Assessments may be conducted by authorities external to the classroom for the purposes of

- Formulating policy.
- Monitoring the effects of policies.
- Enforcing compliance with policies.

- Demonstrating accountability.
- Making comparisons.
- Monitoring progress toward goals.

See Program Standard A

In addition to those purposes, assessments are conducted by school districts to make judgments about the effectiveness of specific programs, schools, and teachers and to report to taxpayers on the district's accomplishments.

The high cost of external assessments and their influence on science teaching practices demand careful planning and implementation. Well-planned, large-scale assessments include teachers during planning and implementation. In addition, all data collected are analyzed, sample sizes are well rationalized, and the sample is representative of the population of interest. This section discusses the characteristics of large-scale assessments.

DATA ANALYSIS

Far too often, more educational data are collected than are analyzed or used to make decisions or take action. Large-scale assessment planners should be able to describe how the data they plan to collect will be used to improve science education.

TEACHER INVOLVEMENT

See Teaching Standard F

The development and interpretation of externally designed assessments for monitoring the educational system should include the active participation of teachers. Teachers' experiences with students make them indispensable participants in the design, development, and interpretation of assessments prepared beyond the classroom. Their involvement helps to ensure congruence of the classroom practice of science education and external assessment practices. Whether at the district, state, or national level, teachers of science need to work with others who make contributions to the assessment process, such as educational researchers, educational measurement specialists, curriculum specialists, and educational policy analysts.

SAMPLE SIZE

The size of the sample on which data are collected depends on the purpose of the assessment and the number of students, teachers, schools, districts, or states that the assessment plan addresses. If, for instance, a state conducts an assessment to learn about student science achievement in comparison with students in another state, it is sufficient to obtain data from a scientifically defined sample of the students in the state. If, however, the purpose of the assessment is to give state-level credit to individual students for science courses, then data must be collected for every student.

REPRESENTATIVE SAMPLE

For all large-scale assessments, even those at the district level, the information should be collected in ways that minimize the time demands on individual students. For many accountability purposes, a sampling design can be employed that has different representative samples of students receiving different sets of tasks. This permits many different dimensions of the science education system to be monitored. Policy makers and taxpayers can make valid inferences about student achievement and opportunity to learn across the nation, state, or district without requiring extensive time commitments from every student in the sample.

Sample Assessments of Student Science Achievement

To illustrate the assessment standards, two examples are provided below. The content standards are stated in terms of understandings and abilities; therefore, the first example is about understanding the natural world. This example requires a body of scientific knowledge and the competence to reason with that information to make predictions, to develop explanations, and to act in scientifically rational ways. The example focuses on predictions and justifying those predictions. The second example is about the ability to inquire, which also requires a body of scientific information and the competence to reason with it to conceptualize, plan, and perform investigations. (These assessment tasks and the content standards do not have a one-to-one correspondence.)

ASSESSING UNDERSTANDING OF THE NATURAL WORLD

The content standards call for scientific understanding of the natural world. Such understanding requires knowing concepts, principles, laws, and theories of the physical, life, and earth sciences, as well as ideas that are common across the natural sciences. That understanding includes the capacity to reason with knowledge. Discerning what a student knows or how the student reasons is not possible without communication, either verbal or representational, a third essential component of understanding.

Inferences about students' understanding can be based on the analysis of their performances in the science classroom and their work products. Types of performances include making class or public presentations, discussing science matters with peers or teachers, and conducting laboratory work. Products of student work include examinations, journal notes, written reports, diagrams, data sets, physical and mathematical models, and collections of natural objects. Communication is fundamental to both performance and product-based assessments.

Clearly relating assessment tasks and products of student work to the valued goals of science education is integral to assessment plans.

Understanding takes different perspectives and is displayed at different levels of sophistication. A physicist's understanding of respiration might be quite different from that of a chemist, just as a cell biologist's understanding of respiration is quite different from that of a physician. The physicist, the chemist, the biologist, and the physician all have a highly sophisticated understanding of respiration. They bring many of the same scientific principles to bear on the concept. However, each is likely to give greater emphasis to concepts that have special significance in their particular discipline. A physicist's emphasis might be on energetics, with little emphasis on the organisms in which respiration takes place. The physician, on the other hand, might emphasize respiration as it specifically applies to humans. The context of application also contributes to differences in per-

spective. The cell biologist's understanding might focus on the mechanisms by which respiration occurs in the cell; the physician's understanding might focus on human respiratory disorders and physical and pathological causes.

An ordinary citizen's understanding of respiration will be much less sophisticated

Eliciting and analyzing explanations are useful ways of assessing science achievement.

than that of a practicing scientist. However, even the citizen's understanding will have different perspectives, reflecting differences in experience and exposure to science.

Legitimate differences in perspectives and sophistication of understanding also will be evident in each student's scientific understanding of the natural world.A challenge to teachers and others responsible for assessing understanding is to decide how such variability is translated into judgments about the degree to which individual students or groups of them understand the natural world. The example that follows illustrates how explanations of the natural world can be a rich source of information about how students understand it.

Because explanation is central to the scientific enterprise, eliciting and analyzing explanations are useful ways of assessing science achievement. The example illustrates how thoughtfully designed assessment exercises requiring explanations provide students with the opportunity to demonstrate the full range of their scientific understanding. Exercises of this sort are not designed to learn whether a student knows a particular fact or concept, but rather to tap the depth and breadth of the student's understanding. Exercises of this sort are difficult to design and are a challenge to score. The example that follows illustrates these challenges.

THE PROMPT. The assessment task begins with a prompt that includes a description of the task and directions. The prompt reads

Some moist soil is placed inside a clear glass jar. A healthy green plant is planted in the soil.The cover is screwed on tightly. The jar is located in a window where it receives sunlight. Its temperature is maintained between 60° and 80°F. How long do you predict the plant will live? Write a justification supporting your prediction. Use relevant ideas from the life ,physical, and earth sciences to make a prediction and justification. If you are unsure of a prediction, your justification should state that, and tell what information you would need to make a better prediction. You should know that there is not a single correct prediction.

Many attributes make the "plant in a jar" a good exercise for assessing understanding. The situation, a plant in a closed jar, can be described to students verbally, with a diagram, or with the actual materials, thus eliminating reading as a barrier to a student response. The situation can be understood by students of all ages, minimizing students' prior knowledge of the situation as a factor in ability to respond. The explanation for the prediction can be developed at many different levels of complexity, it can be qualitative or quantitative. It can be based on experience or theory, and it uses ideas from the physical,life,and earth sciences, as well as cross-disciplinary ideas, thus allowing students to demonstrate the full range of

See Teaching Standard B

their understanding of science at various levels of their study of science.

DEVELOPING SCORING RUBRICS. The process of scoring student-generated explanations requires the development of a scoring rubric. The rubric is a standard of performance for a defined population. Typically, scoring rubrics are developed by the teachers of the students in the target population. The performance standard is developed through a consensus process called "social moderation." The steps in designing a scoring rubric involve defining the performance standard for the scientifically literate adult and then deciding which elements of that standard are appropriate for students in the target population. The draft performance standard is refined by subsequent use with student performance and work. Finally, student performances with respect to the rubric are differentiated. Performances are rated satisfactory, exemplary, or inadequate. Differences in opinions about the rubric and judgments about the quality of students' responses are moderated by a group of teachers until consensus is reached for the rubric.

Because a target population has not been identified,and rubrics need to function in the communities that develop them, this section does not define a rubric. Rather the steps in developing a rubric are described.

THE PERFORMANCE OF A SCIENTIFICALLY LITERATE ADULT. Developing a scoring rubric begins with a description of the performance standard for scientifically literate adults. That performance standard is developed by a rubric development team. Members of the team write individual responses to the exercise that reflect how each believes a scientifically literate adult should respond. They also seek responses from other adults. Based on the individual responses, the team negotiates a team response that serves as the initial standard.

The team's standard is analyzed into the components of the response. In the plant-in-a-jar exercise, the components are the predictions, the information used to justify the predictions, the reasoning used to justify predictions, and the quality of the communication.

Examples of predictions from a scientifically literate adult about how long the plant can live in the jar might include (1) insufficient information to make a prediction, (2) the plant can live indefinitely, or (3) insects or disease might kill the plant. Whatever the prediction, it should be justified. For example, if the assertion is made that the information provided in the prompt is insufficient to make a prediction, then the explanation should describe what information is needed to make a prediction and how that information would be used.

Scientifically literate adults will rely on a range of knowledge to justify their predictions. The standard response developed by the team of teachers will include concepts from the physical, life, and earth sciences, as well as unifying concepts in science. All are applicable to making and justifying a prediction about the life of a plant in a jar, but because of the differences in emphasis, no one person would be expected to use all of them. Some concepts, such as evaporation, condensation, energy (including heat, light, chemical), energy conversions, energy transmission, chemical interactions, catalysis, conservation of mass,and dynamic equilib-

See Content Standards B, C,and D (all grade levels)

rium,are from physical science. Other concepts are from life sciences, such as plant physiology, plant growth, photosynthesis, respiration, plant diseases, and plant and insect interaction.Still other concepts are from the earth sciences, such as soil types, composition of the atmosphere, water cycle, solar energy, and mineral cycle. Finally, unifying concepts might be used to predict and justify a prediction about the plant in the jar. Those might include closed, open, and isolated systems; physical models; patterns of change; conservation; and equilibrium.

See Unifying Concepts and Processes in Chapter 6

The knowledge required to predict the life of a plant in a jar is not be limited to single concepts.A deeper understanding of the phenomena could be implied by a justification that includes knowledge of chemical species and energy. Keeping track of energy, of $C_6H_{12}O_6$ (sugar), CO_2 (carbon dioxide), H_2O (water), and O_2 (oxygen), and of minerals requires knowing about the changes they undergo in the jar and about equilibria among zones in the jar (soil and atmosphere). The jar and its contents form a closed system with respect to matter but an open system with respect to energy. The analysis of the life expectancy of the plant in the jar also requires knowing that the matter in the jar changes form, but the mass remains constant. In addition, knowing that gases from the atmosphere and minerals in the soil become a part of the plant is important to the explanation.

A well-crafted justification . . . demonstrates reasoning characterized by a succession of statements that follow one another logically without gaps from statement to statement.

A deeper understanding of science might be inferred from a prediction and justification that included knowledge of the physical chemistry of photosynthesis and respiration. Photosynthesis is a process in which radiant energy of visible light is converted into chemical bond energy in the form of special carrier molecules, such as ATP, which in turn are used to store chemical bond energy in carbohydrates. The process begins with light absorption by chlorophyll, a pigment that gives plants their green color. In photosynthesis, light energy is used to drive the reaction:

Carbon dioxide + water → sugar + oxygen.

Respiration is a process in which energy is released when chemical compounds react with oxygen. In respiration, sugars are broken down to produce useful chemical energy for the plant in the reaction:

Sugar + oxygen → water + carbon dioxide.

Photosynthesis and respiration are complementary processes, because photosynthesis results in the storage of energy, and respiration releases it. Photosynthesis removes CO_2 from the atmosphere; respiration adds CO_2 to the atmosphere.

A justification for a prediction about the life of the plant in the jar might include knowledge of dynamic equilibrium. Equilibrium exists between the liquid and vapor states of water. The liquid water evaporates continuously. In the closed container, at constant temperature,the rate of condensation equals the rate of evaporation. The water is in a state of dynamic equilibrium.

Another attribute of a well-crafted justification relates to assumptions. The justification should be explicit about the assumptions that underlie it and even contain some speculation concerning the implications of making alternative assumptions.

Finally, a well-crafted justification for any prediction about the plant in the jar demonstrates reasoning characterized by a succession of statements that follow one another logically without gaps from statement to statement.

SCORING RUBRICS FOR DIFFERENT POPULATIONS OF STUDENTS. The plant-in-a-jar assessment exercise is an appropriate prompt for understanding plants at any grade level. Development of the scoring rubrics for students at different grade levels requires consideration of the science experiences and developmental level of the students. For instance,the justifications of students in elementary school could be expected to be based primarily on experiences with plants. Student justifications would contain little, if any, scientific terminology. A fourth-grade student might respond to the exercise in the following way:

The plant could live. It has water and sunlight. It could die if it got frozen or a bug eats it. We planted seeds in third grade. Some kids forgot to water them and they died. Eddie got scared that his seeds would not grow. He hid them in his desk.They did. The leaves were yellow. After Eddie put it in the sun it got green.The plants in our terrarium live all year long.

Expectations for justifications constructed by students in grades 5-8 are different. These should contain more generalized knowledge and use more sophisticated language and scientific concepts such as light, heat, oxygen, carbon dioxide, energy, and photosynthesis.

By grade 12, the level of sophistication should be much higher. Ideally, the 12th grader would see the plant in a jar as a physical model of the Earth's ecosystem, and view photosynthesis and respiration as complementary processes.

Setting a performance standard for a population of students depends on the population's developmental level and their experiences with science. Considerations to be made in using student responses for developing a rubric can be illustrated by discussing two justifications constructed by students who have just completed high-school biology. Student E has constructed an exemplary justification for her prediction about the plant in the jar. Student S has constructed a less satisfactory response but has not completely missed the point.

STUDENT E: *If there are no insects in the jar or microorganisms that might cause some plant disease,the plant might grow a bit and live for quite a while. I know that when I was in elementary school we did this experiment. My plant died—it got covered with black mold. But some of the plants other kids had got bigger and lived for more than a year.*

The plant can live because it gets energy from the sunlight.When light shines on the leaves, photosynthesis takes place. Carbon dioxide and water form carbohydrates and oxygen. This reaction transforms energy from the sun into chemical energy. Plants can do this because they have chlorophyll.

The plant needs carbohydrates for life processes like growing and moving. It uses the carbohydrates and oxygen to produce energy for life processes like growth and motion. Carbon dioxide is produced too.

After some time the plant probably will stop growing. I think that happens when all the

minerals in the soil are used up. For the plant to grow it needs minerals from the soil. When parts of the plant die, the plant material rots and minerals go back into the soil. So that's why I think that how much the plant will grow will depend on the minerals in the soil.

The gases, oxygen, carbon dioxide and water vapor just keep getting used over and over. What I'm not sure about is if the gases get used up. Can the plant live if there is no carbon dioxide left for photosynthesis? If there is no carbon dioxide, will the plant respire and keep living?

I'm pretty sure a plant can live for a long time sealed up in a jar, but I'm not sure how long or exactly what would make it die.

STUDENT S: *I believe that putting a small plant in a closed mayonnaise jar at 60-80°F is murder. I believe that this plant will not last past a week (3 days). This is so for many reasons. Contained in a jar with constant sunlight at 80°F the moisture in the soil will most likely start to evaporate almost immediately. This will leave the soil dry while the air is humid. Since we are in a closed container no water can be restored to the soil (condensation). This in turn will cause no nutrients from the soil to reach the upper plant, no root pressure!*

Besides this, with photosynthesis occurring in the leaves, at least for a short time while water supplies last, the CO_2 in the air is being used up and O_2 is replacing it. With no CO_2 and no H_2O, no light reaction and/or dark reaction can occur and the plant can't make carbohydrates. The carbohydrates are needed for energy.

In condusion, in a jar closed from CO_2 and water, plants use up their resources quickly, preventing the equation $CO_2 + H_2O \rightarrow C_6H_{12}O_6$ and O_2 and, therefore, energy from carbohydrates.

This jar also works as a catalyst to speed up the process by causing evaporation of H_2O through incomplete vaporization. This would shut down the root hair pressure in the plant which allows water (if any) + nutrients to reach the leaves. All in all the plant will not live long (3 days at the most then downhill).

Judging the quality of information contained in a justification requires consensus on the information contained in it and then using certain standards to compare that information with the information in the rubric. Standards that might be applied include the scientific accuracy of the information in the justification, the appropriateness of the knowledge to the student's age and experience, the sophistication of the knowledge, and the appropriateness of the application of the knowledge to the situation.

Foremost, judgments about the quality of the information contained in the justification should take into account the accuracy of the information the student used in crafting the response. Student S's justification contains some misinformation about the water evaporation-condensation cycle and about dynamic equilibrium in closed systems. The statements in which this inference is made are "the moisture in the soil will most likely start to evaporate almost immediately. This will leave the soil dry while the air is humid. Since we are in a closed container no water can be restored to the soil (condensation)." The student's statement that "soil in the container will be dry while the air is humid," suggests lack of knowledge about equilibrium in a closed system. In contrast with Student S's misinformation, Student E's justification contains information that is neither unusually sophisticated when viewed against the content of most high-school biology texts, nor erroneous.

Judgments about the appropriateness of the information are more difficult to make. A person familiar with the biology course

the student took might assert that the information in the student's response is not as sophisticated as what was taught in the course. In that case, the content of the biology course is being used as the standard for rating the quality of the responses. Alternatively, the standard for rating the appropriateness of the information might be the scientific ideas in the content standards.

Student E's response is rated higher than Student S's on the basis of the quality of information.Student S's justification provides some information about what the student does and does not know and provides some evidence for making inferences about the structure of the student's knowledge. For example,the student did not consider the complementary relationship of photosynthesis and respiration in crafting the justification. Perhaps the student does not know about respiration or that the processes are complementary. Alternatively, the two concepts may be stored in memory in a way that did not facilitate bringing both to bear on the exercise. Testing the plausibility of the inferences about the student's knowledge structure would require having a conversation with the student. To learn if the student knows about respiration, one simply has to ask. If the student knows about it and did not apply it in making the prediction, this is evidence that respiration is not understood in the context of the life processes of plants.

Student E's response is well structured and consistent with the prediction. The statements form a connected progression. The prediction is tentative and the justification indicates it is tentative due to the lack of information in the prompt and the student's uncertainty about the quantitative details of the condition under which photosynthesis and respiration occur. The student is explicit about certain assumptions, for instance, the relationship of minerals in the soil and plant growth. The questions the student poses in the justification can be interpreted as evidence that alternative assumptions have been considered.

In contrast, Student S's prediction is stated with unwarranted assurance and justified without consideration of anything more than the availability of sufficient water. Furthermore, the justification does not proceed in a sequential way, proceeding from general principles or empirical evidence to a justification for the prediction.

Student S's response highlights an important point that justifies separating the scoring of information from the scoring of reasoning. A student can compose a well-reasoned justification using incorrect information. For instance, had the student posed the following justification, the reasoning would be adequate even if the conclusion that the soil is dry were not correct. The reasoning would be rated higher had the student communicated that

Plants need energy to live. Plants get energy from sunlight through a process of photosynthesis. Plants need water to photosynthesize. Because the soil is dry, water can't get to the leaves, the plant can't photosynthesize and will die from lack of energy.

Developing scoring rubrics through moderation requires highly informed teachers experienced in the process. Even when teachers are adequately prepared, the moderation process takes time. The content standards call for knowledge with understanding.

Considerable resources must therefore be devoted to preparing teachers and others in the science education system to design and rate assessments that require students to display understanding, such as just described.

ASSESSING THE ABILITY TO INQUIRE

The second assessment example focuses on inquiry. The content standards call for understanding scientific inquiry and developing the ability to inquire. As in understanding the natural world, understanding and doing inquiry are contingent on knowing concepts, principles,laws, and theories of the physical, life, and earth sciences. Inquiry also requires reasoning capabilities and skills in manipulating laboratory or field equipment.

As in understanding the natural world, inferences about students' ability to inquire and their understanding of the process can be based on the analysis of performance in the science classroom and work products.

Understanding and doing inquiry are contingent on knowing concepts, principles, laws, and theories of the physical, life, and earth sciences.

The example that follows describes twelfth grade students' participation in an extended inquiry. The exercise serves two purposes. It provides the teacher with information about how well students have met the inquiry standards. Equally important, it serves as a capstone experience for the school science program. The extended inquiry is introduced early in the school year. It involves students working as individuals and in small groups investigating a question of their choice.

IDENTIFYING A WORTHWHILE AND RESEARCHABLE QUESTION. Throughout the school science program, students have been encouraged to identify questions that interest them and are amenable to investigation. These questions are recorded in student research notebooks. Early in the senior year of high school,students prepare draft statements of the question they propose to investigate and discuss why that question is a reasonable one. Those drafts are circulated to all members of the class. Students prepare written reviews of their classmate's proposals, commenting on the quality of the research question and the rationale for investigating it. Students then revise their research question based on peer feedback. Finally, students present and defend their revised questions to the class.

PLANNING THE INVESTIGATION. The teacher encourages but does not require students to work together in research groups of two to four students. After presenting research questions to the class,students form the research groups, which come to agreement on a question to investigate and begin developing a preliminary plan for conducting the investigation. Each individual in the group is required to keep extensive records of the group's work, especially documenting the evolution of their final research question from the several questions originally proposed. As plans for investigations evolve, the research questions are sharpened and modified to meet the practical constraints of time and resources avail-

able to the class. Each student maintains journal notes on this process. When a group is satisfied that their plan has progressed to the point where work can begin, the plan is presented to classmates. Written copies of the plan are distributed for written review, followed by a class seminar to discuss each research plan.On the basis of peer feedback, each group revises its research plan, recognizing that as the plan is implemented, it will require still further revisions. Each student in the class is responsible for reviewing the research plan of every group, including a written critique and recommendations for modifying the plan.

EXECUTING THE RESEARCH PLAN. During this phase of the extended investigation, students engage in an iterative process involving assembling and testing apparatus; designing and testing forms of data collection; developing and testing a data collection schedule; and collecting, organizing, and interpreting data.

DRAFTING THE RESEARCH REPORT. Based on the notes of individuals, the group prepares a written report, describing the research. That report also includes data that have been collected and preliminary analysis. Based on peer feedback, the groups modify their procedures and continue data collection. When a group is convinced that the data-collection method is working and the data are reasonably consistent,they analyze the data and draw conclusions.

After a seminar at which the research group presents its data, the analysis,and conclusions,the group prepares a first draft of the research report. This draft is circulated to classmates for preparation of individual critiques. This feedback is used by the group to prepare its final report.

ASSESSING INDIVIDUAL STUDENT ACHIEVEMENT. While the class is engaged in the extended investigation, the teacher observes each student's performance as the student makes presentations to the class, interacts with peers, and uses computers and laboratory apparatus. In addition, the teacher has products of the individual student's work, as well as group work, including draft research questions, critiques of other student work, and the individual student's research notebook. Those observations of student performance and work products are a rich source of data from which the teacher can make inferences about each student's understanding of scientific ideas and the nature of scientific inquiry. For instance, in the context of planning the inquiry, students pose questions for investigation. Their justifications for why the question is a scientific one provide evidence from which to infer the extent and quality of their understanding of the nature of science, understanding of the natural world, understanding of the life, physical, and earth sciences, as well as the quality and extent of their scientific knowledge and their capacity to reason scientifically.

Evidence for the quality of a student's ability to reason scientifically comes from the rationale for the student's own research question and from the line of reasoning used to progress from patterns in the collected data to the conclusions. In the first instance, the student distills a research question from an understanding of scientifi c

ideas associated with some natural phenomenon. In the second instance, the student generates scientific information based on data. In either case, the quality of the reasoning can be inferred from how well connected the chain of reasoning is, how explicit the student is about the assumptions made, and the extent to which speculations on the implications of having made alternative assumptions are made.

The writing and speaking requirements of this extended investigation provide ample evidence for assessing the ability of the student to communicate scientific ideas.

CHANGING EMPHASES

The *National Science Education Standards* envision change throughout the system. The assessment standards encompass the following changes in emphases:

LESS EMPHASIS ON	MORE EMPHASIS ON
Assessing what is easily measured	Assessing what is most highly valued
Assessing discrete knowledge	Assessing rich, well-structured knowledge
Assessing scientific knowledge	Assessing scientific understanding and reasoning
Assessing to learn what students do not know	Assessing to learn what students do understand
Assessing only achievement	Assessing achievement and opportunity to learn
End of term assessments by teachers	Students engaged in ongoing assessment of their work and that of others
Development of external assessments by measurement experts alone	Teachers involved in the development of external assessments

References for Further Reading

Baron, J.B. 1992.SEA Usage of Alternative Assessment: The Connecticut Experience. In Focus on Evaluation and Measurement, vol. 1 and 2. Proceedings of the National Research Symposium on Limited English Proficient Student Issues.. Washington, DC.

Baxter, G.P.,R.J. Shavelson, and J. Pine. 1992. Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement,29 (1): 1-17.

Champagne,A.B.,and S.T. Newell.1992. Directions for Research and Development: Alternative Methods of Assessing Scientific Literacy. Journal of Research in Science Teaching 29 (8):841-860.

Glaser, R.1992. Cognitive Theory as the Basis for Design of Innovative Assessment: Design Characteristics of Science Assessments. Los Angeles,CA: National Center for Research on Evaluation, Standards,and Student Testing.

Koretz, D., B. Stecher, S. Klein, and D. McCaffrey. 1994. The Vermont Portfolio Assessment Program: Findings and Implications. Educational Measurement: Issues and Practice,13 (3): 5-16.

Loucks-Horsley, S.,R. Kapitan,M.O. Carlson, P.J. Kuerbis,R.C. Clark, G.M. Nelle, T.P. Sachse, and E. Walton.1993. Elementary School Science for the '90s. Andover, MA: The Network, Inc.

Messick, S.1994. The interplay of evidence and consequences in the validation of performance assessments.Educational Researcher, 23 (2): 13-23.

Moss, P.A. 1994. Can there be validity without reliability? Educational Researcher, 23 (2): 13-23.

NRC (National Research Council).1991. Improving Instruction and Assessment in Early Childhood Education: Summary of a Workshop Series. Washington, DC: National Academy Press.

NRC (National Research Council).1989. Fairness in Employment Testing: Validity Generalization, Minority Issues,and the General Aptitude Test Battery. J.A. Hartigan, and A.K. Wigdor, eds. Washington, DC: National Academy Press.

Oakes, J., T. Ormseth,R. Bell,and P. Camp. 1990. Multiplying Inequalities: The Effect of Race, Social Class,and Tracking on Students' Opportunities to Learn Mathematics and Science. Santa Monica, CA: RAND Corporation.

Raizen, S.A., J.B. Baron,A.B. Champagne,E. Haertel,I.V. Mullis,and J. Oakes. 1990. Assessment in Science Education: The Middle Years. Washington, DC: National Center for Improving Science Education.

Raizen, S.A., J.B. Baron,A.B. Champagne,E. Haertel,I.V. Mullis,and J. Oakes. 1989. Assessment in Elementary School Science Education. Washington, DC: National Center for Improving Science Education.

Ruiz-Primo, M.A.,G.P. Baxter, and R.J. Shavelson.1993.On the stability of performance assessments. Journal of Educational Measurement,30 (1): 41-53.

Shavelson,R.J. 1991. Performance assessment in science. Applied Measurement in Education, 4 (4):347-62.

Shavelson, R.J.,G. Baxter, and J. Pine. 1992. Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21 (4): 22-27.