Chapter 13

# Assessing for Learning: Performance Assessment

This chapter will help you answer the following questions about your learners:

- Can complex cognitive outcomes, such as critical thinking and decision making, be more effectively learned with performance tests than with traditional methods of testing?
- How can I construct performance tests that measure self-direction, ability to work with others, and social awareness?
- Can standardized performance tests be scored reliably?
- How do I decide what a performance test should measure?
- How do I design a performance test based on real-life problems important to people who are working in the field?
- How can I use a simple checklist to score a performance test accurately and reliably?
- What are some ways of using rating scales to score a performance test?
- How do I decide how many total points to assign to a performance test?
- How do I decide what conditions to place on my learners when completing a performance test to make it as authentic as possible?
- What are student portfolios and how can they be graded fairly and objectively?

- How do I weight performance tests and combine them with other student work, such as quizzes, homework, and class participation, to create a final grade?

In this chapter you will also learn the meanings of these terms:

**authentic assessment**
**holistic scoring**
**multimodal assessment**
**performance testing**
**portfolio assessment**
**primary trait scoring**
**rubrics**
**testing constraints**

Lori Freeman, the chair of the seventh-grade science department at Sierra Blanca Junior High School, is holding a planning session with her science teachers. The topic is evaluation of the seventh-grade life science course. Ms. Freeman had previously assigned several faculty to a committee to explore alternatives to multiple-choice tests for assessing what seventh-graders achieved after a year of life science, so she begins this second meeting with a summary of the decisions made by the committee.

**Ms. Freeman:** Recall that last time we decided to try performance assessment on a limited basis. To begin, we decided to build a performance assessment for our unit on photosynthesis. Does anyone have anything to add before we get started?
**Ms. Brown:** I think it’s important that we look at different ways our students can demonstrate that they can do science rather than just answer multiple-choice and essay questions. But I also want to make sure we’re realistic about what we’re getting into. I have 150 students in seventh-grade life science. From what I heard

last time, a good performance assessment is very time-consuming. I don't see how we can make every test performance based.

Ms. Freeman: Nor should we. Paper-and-pencil tests will always be a principal means of assessment, but I think we can measure reasoning skills, problem solving, and critical thinking better than we're doing now.

**Mr. Hollyfield:** And recognize that there are a variety of ways students can show they're learning science. Right now there's only one way—a multiple-choice test.

**Mr. Moreno:** I think Jan's concerns are real. We have to recognize that performance assessment takes a lot of time. But don't forget that a good performance assessment, basically, is a good lesson. A lot of performance testing is recording what learners are doing during the lesson. We just have to do it in a more systematic way.

**Ms. Ellison:** I'm concerned about the subjectivity of these types of assessments. From what I know, a lot of performance assessment is based on our own personal judgment or rating of what students do. I'm not sure I want to defend to a parent a low grade that is based on my personal feelings.

**Ms. Freeman:** That can be a problem. Remember, though, you make some subjective judgments now when you grade essays or select the multiple-choice questions you use. And as with paper-and-pencil tests, there are ways to score performance assessments objectively and reliably. I think knowing that all our learners will have to demonstrate skills in critical thinking, problem solving, and reasoning will make us do a better job of teaching. I know we shouldn't let tests dictate what we teach. But in the case of performance assessment, maybe that's not such a bad idea.

What exactly is performance assessment? What form does it take? How, when, and why is it used? What role does performance assessment have in conjunction with more traditional forms of assessment? How does a teacher acquire proficiency in designing

and scoring performance tests? In this chapter, we will introduce you to performance assessment. First we will describe performance assessment by showing examples of performance tests currently being used in elementary and secondary schools. We will show the progress educators have made at the state and national levels in developing performance tests that are objective, practical, and efficient. Then we will show you how to start developing and using performance tests in your own classroom.

## Performance Testing

In Chapters 4, 5, and 6, you learned that children acquire a variety of skills in school. Some of them require learners to take in information by memorizing vocabulary, multiplication tables, dates of historical events, and so on. Other skills involve learning action sequences or procedures to follow when performing mathematical computations, dissecting a frog, focusing a microscope, writing, or typing. In addition, you learned that students must acquire concepts, rules, and generalizations that allow them to understand what they read, analyze and solve problems, carry out experiments, write poems and essays, and design projects to study historical, political, or economic problems.

Some of these skills are best assessed with paper-and-pencil tests. But other skills—particularly those involving independent judgment, critical thinking, and decision making—are best assessed with **performance testing.** Although paper-and-pencil tests currently represent the principal means of assessing these more complex cognitive outcomes, in this chapter we will study other ways of measuring them in more authentic contexts.

## Performance Tests: Direct Measures of Competence

In Chapters 11 and 12 you learned that many psychological and educational tests measure learning indirectly. That is, they ask questions whose responses indicate that

something has been learned or mastered. Performance tests, on the other hand, use direct measures of learning rather than indicators that simply suggest cognitive, affective, or psychomotor processes have taken place. In athletics, diving and gymnastics are examples of performances that judges rate directly. Likewise, at band contests judges directly see and hear the competence of a trombone or tuba player and pool their ratings to decide who makes the state or district band and who gets the leading chairs.

Teachers can use performance tests to assess complex cognitive learning as well as attitudes and social skills in academic areas such as science, social studies, or math. When doing so, you establish situations that allow you to directly observe and rate learners as they analyze, problem solve, experiment, make decisions, measure, cooperate with others, present orally, or produce a product. These situations simulate real-world activities that students might be expected to perform in a job, in the community, or in various forms of advanced training.

Performance tests also allow teachers to observe achievements, habits of mind, ways of working, and behaviors of value in the real world. In many cases, these are outcomes that conventional tests may miss. Performance tests can include observing and rating learners as they carry out a dialogue in a foreign language, conduct a science experiment, edit a composition, present an exhibit, work with a group of other learners to design a student attitude survey, or use equipment. In other words, the teacher observes and evaluates student abilities to carry out complex activities that are used and valued outside the immediate confines of the classroom.

## Performance Tests Can Assess Processes and Products

Performance tests can be assessments of processes, products, or both. For example, at the Darwin School in Winnipeg, Manitoba, teachers assess the reading process of each student by noting the percentage of words read accurately during oral reading, the number of sentences read by the learner that are meaningful within the context of the

story, and the percentage of story elements that the learner can talk about in his or her own words after reading.

At the West Orient school in Gresham, Oregon, fourth-grade learners assemble portfolios of their writing products. These portfolios include both rough and polished drafts of poetry, essays, biographies, and self-reflections. Several math teachers at Twin Peaks Middle School in Poway, California, require their students to assemble math portfolios, which include the following products of their problem-solving efforts: long-term projects, daily notes, journal entries about troublesome test problems, written explanations of how they solved problems, and the problem solutions themselves.

Social studies learning processes and products are assessed in the Aurora, Colorado, public schools by engaging learners in a variety of projects built around this question: "Based on your study of Colorado history, what current issues in Colorado do you believe are the most important to address, what are your ideas about the resolutions of those issues, and what contributions will you make toward the resolutions?" (Pollock, 1992). Learners answer these questions in a variety of ways involving individual and group writing assignments, oral presentations, and exhibits.

## Performance Tests Can Be Embedded in Lessons

The examples of performance tests just given involve performances that occur outside the context of a lesson and are completed at the end of a term or during an examination period. Many teachers use performance tests as part of their lessons. In fact, some proponents of performance tests hold that the ideal performance test is a good teaching activity (Shavelson & Baxter, 1992). Viewed from this perspective, a well-constructed performance test can serve as a teaching activity as well as an assessment.

For example, Figure 13.1 illustrates a performance activity and assessment that was embedded in a unit on electricity in a general science class (Shavelson & Baxter, 1992). During the activity the teacher observes and rates each learner on the method he or she

uses to solve the problem, the care with which he or she measures, the manner of recording results, and the correctness of the final solution. This type of assessment provides immediate feedback on how learners are performing, reinforces hands-on teaching and learning, and underscores for learners the important link between teaching and testing. In this manner, it moves the instruction toward higher-order thinking.

Other examples of lesson-embedded performance tests might include observing and rating the following as they are actually happening: typing, preparing a microscope slide, reading, programming a calculator, giving an oral presentation, determining how plants react to certain substances, designing a questionnaire or survey, solving a math problem, developing an original math problem and a solution for it, critiquing the logic of an editorial, or graphing information.

## Performance Tests Can Assess Affective and Social Skills

Educators across the country are using performance tests to assess not only higher-level cognitive skills but also noncognitive outcomes such as self-direction, ability to work with others, and social awareness (Redding, 1992). This concern for the affective domain of learning reflects an awareness that the skilled performance of complex tasks involves more than the ability to recall information, form concepts, generalize, and problem solve. It also involves habits of mind, attitudes, and social skills.

The Aurora public schools in Colorado have developed a list of learning outcomes and their indicators for learners in grades K through 12. These are shown in Table 13.1. For each of these 19 indicators, a four-category rating scale has been developed to serve as a guide for teachers who are unsure how to define “assumes responsibility” or “demonstrates consideration.” While observing learners during performance tests in social studies, science, art, or economics, teachers recognize and rate those behaviors that suggest learners have acquired the outcomes.

Teachers in Aurora are encouraged to use this list of outcomes when planning their courses. They first ask themselves what content—key facts, concepts, and principles—all learners should remember. In addition, they try to fuse this subject area content with the five district outcomes by designing special performance tests. For example, a third-grade language arts teacher who is planning a writing unit might choose to focus on indicators 8 and 9 to address district outcomes related to "collaborative worker," indicator 1 for the outcome "self-directed learner," and indicator 13 for the outcome "quality producer." She would then design a performance assessment that allows learners to demonstrate learning in these areas. She might select other indicators and outcomes for subsequent units and performance tests.

Likewise, a ninth-grade history teacher, having identified the important content for a unit on civil rights, might develop a performance test to assess district outcomes related to "complex thinker," "collaborative worker," and "community contributor." A performance test (adapted from Redding, 1992, p. 51) might take this form: "A member of a minority in your community has been denied housing, presumably on the basis of race, ethnicity, or religion. What steps do you believe are legally and ethically defensible, and in what order do you believe they should be followed?" This performance test could require extensive research, group collaboration, role-playing, and recommendations for current ways to improve minority rights.

Performance tests represent an addition to the testing practices reviewed in the previous chapter. They are not intended to replace these practices. Paper-and-pencil tests are the most efficient, reliable, and valid instruments available for assessing knowledge, comprehension, and some types of application. But when it comes to assessing complex thinking skills, attitudes, and social skills, properly constructed performance tests can do a better job. On the other hand, if they are not properly constructed, performance assessments can have some of the same problems with scoring efficiency, reliability, and validity as traditional approaches to testing. In this chapter, we will guide you through a

process that will allow you to properly construct performance tests in your classroom. Before doing so, let's look at what educators and psychologists are doing at the national and state levels to develop standardized performance tests. As you read about these efforts, pay particular attention to how these forms of standardized assessment respond to the needs for scoring efficiency, reliability, and validity.

## Standardized Performance Tests

In the past decade several important developments have highlighted concerns about our academic expectations for learners and how we measure them. Several presidential commissions completed comprehensive studies of the state of education in American elementary and secondary schools (Goodlad, 1984; Holmes Group, 1990; Sizer, 1984). They concluded that instruction at all levels is predominantly focused on memorization, drills, and workbook exercises. They called for the development of a curriculum that focuses on teaching learners how to think critically, reason, and problem solve in real-world contexts.

Four teachers' organizations—the National Council of Teachers of Mathematics, the National Council for Social Studies, the National Council for Improving Science Education, and the National Council of Teachers of English—took up the challenge of these commissions by publishing new curriculum frameworks. These frameworks advocate that American schools adopt a "thinking curriculum" (Mitchell, 1992; Parker, 1991; Willoughby, 1990). Finally, the National Governors' Association (NGA) in 1990 announced six national goals for American education, two of which target academic achievement:

> Goal 3: American students will achieve competency in English, mathematics, science, history, and geography at grades 4, 8, and 12 and will be prepared for

responsible citizenship, further learning, and productive employment in a modern economy.

Goal 4: U.S. students will be the first in the world in science and mathematics achievement.

The NGA commissioned the National Educational Goals Panel to prepare an annual report on the progress made by American schools toward achieving these goals. Its first report, in September 1991, concluded that since no national examination system existed, valid information could not be gathered on the extent to which American schools were accomplishing Goals 3 and 4. The goals panel then set up two advisory groups to look into the development of a national examination system: the National Council on Education Standards and Testing and the New Standards Project. Both groups concluded that Goals 3 and 4 would not be achieved without the development of a national examination system to aid schools in focusing their curricula on critical thinking, reasoning, and problem solving. Moreover, these groups agreed that only a performance-based examination system would adequately accomplish the task of focusing schools on complex cognitive skills.

The challenge for these groups was to overcome the formidable difficulties involved in developing a standardized performance test. At a minimum, such tests must have scoring standards that allow different raters to compute similar scores regardless of when or where the scoring is done. How then does the New Standards Project propose to develop direct measures of learning in science, mathematics, and the social studies with national or statewide standards that all schools can measure reliably?

To help in this process, several states, including California, Arizona, Maryland, Vermont, and New York, have developed standardized performance tests in the areas of writing, mathematics, and science. They have also worked out procedures to achieve scoring reliability of their tests. For example, New York's Elementary Science Performance Evaluation Test (ESPET) was developed over a number of years with the

explicit purpose of changing how teachers taught and students learned science (Mitchell, 1992). It was first administered on a large scale in 1989. Nearly 200,000 fourth-grade students in 4,000 of New York's public and nonpublic schools took the ESPET. They included students with learning disabilities and physical handicaps as well as other learners traditionally excluded from such assessment. The fact that all fourth-grade learners took this test was intended to make a statement that science education and complex learning are expected of all students.

ESPET contains seven sections. Some contain more traditional multiple-choice and short-essay questions, and others are more performance based. Following is a description of the manipulative skills section of the test:

> Five balances are seemingly randomly distributed across the five rows and five columns of desks. The balances are obviously homemade: the shaft is a dowel; the beam is fixed to it with a large nail across a notch; and the baskets, two ordinary plastic salad bowls, are suspended by paper clips bent over the ends of the beams. Lumps of modeling clay insure the balance. On the desk next to the balance beam are a green plastic cup containing water, a clear plastic glass with a line around it halfway up, a plastic measuring jug, a thermometer, and ten shiny new pennies.
>
> Other desks hold electric batteries connected to tiny light bulbs, with wires running from the bulbs ending in alligator clips. Next to them are plastic bags containing spoons and paper clips. A single box sits on other desks. Another desk holds pieces of paper marked A, B, and C, and a paper container of water. The last setup is a simple paper plate divided into three parts for a TV dinner, with labeled instructions and a plastic bag containing a collection of beans, peas, and corn....
>
> Children silently sit at the desks absorbed in problem solving. One boy begins the electrical test, makes the bulb light, and lets out a muffled cry of satisfaction. The instructions tell him to test the objects in the plastic bag to see if they can

make the bulb light. He takes the wire from one of the plastic bags and fastens an alligator clip to it. Nothing happens and he records a check in the “bulb does not light” column on his answer sheet. He gets the same result from the toothpick. He repeats the pattern for all five objects....

Meanwhile, in the same row of desks, a girl has dumped out the beans and peas into the large division of the TV dinner plate as instructed and is classifying them by placing them into the two smaller divisions. She puts the Lima beans and the kidney beans into one group and the pintos, peas, and corn into the other group. The first group, she writes, is “big and dull”; the second is “small and colorful.”

At the end of seven minutes, the teacher instructs them to change desks. Every child must rotate through each of the five science stations. In one day, the school tests four classes each with about twenty-five children. One teacher is assigned to set up and run the tests. The classroom teachers bring in their classes at intervals of about one hour. (Mitchell, 1992)

## A Test Worth Studying For

The ESPET is a syllabus-driven performance examination. In other words, its development began with the creation of a syllabus: a detailed specification of the content and skills on which learners will be examined and the behaviors that are accepted as indicators that the content and skills have been mastered. A syllabus does not specify how the content and skills will be taught. These details, which include specific objectives, lesson plans, and activities, are left to the judgment of the teacher. The syllabus lets the teacher (and learner) know what is on the exam by identifying the real-world behaviors, called *performance objectives*, learners must be able to perform in advanced courses, other programs of study, or in a job.

Teachers of different grades can prepare learners for these objectives in numerous ways, a preparation that is expected to take several years. The examination and the

syllabus and performance objectives that drive it are a constant reminder to learners, parents, and teachers of the achievements that are to be the end products of their efforts. Tests like the ESPET, by virtue of specifically defining the performances to be achieved, represent an **authentic assessment** of what is taught.

## A Test Worth Teaching To

But if teachers know what's on the ESPET, won't they narrow their teaching to include only those skills and activities that prepare students for the exam? Performance test advocates, such as Resnick (1990) and Mitchell (1992), argue that teaching to a test has not been a concern when the test involves gymnastics, diving, piano playing, cooking, or repairing a radio. This is because these performances are not solely test-taking tasks but also job and life tasks necessary for adult living. Performance tests, if developed correctly, should also include such tasks. Here is Ruth Mitchell's description of a worst-case scenario involving teaching to the ESPET:

> Suppose as the worst case (and it is unlikely to happen) that a Grade 4 teacher in New York State decides that the students' scores on the manipulative skills test next year will be perfect. The teacher constructs the whole apparatus as it appeared in the test classroom...and copies bootlegged answer sheets. And, suppose the students are drilled on the test items, time after time. By the time they take the test, these students will be able to read and understand the instructions. They will know what "property" means in the question, "What is another property of an object in the box?" (This word was the least known of the carefully chosen vocabulary in 1989.) The students will be able to write comprehensible answers on the answer sheets. Further, they will have acquired extremely important skills in using measuring instruments, predicting, inferring, observing, and classifying. In teaching as opposed to a testing situation, it will become clear that there is no right answer to a classification, only the development of a defensible

> criterion....In every case, the students' manipulative skills will be developed along with their conceptual understanding.
>
> A class that did nothing beyond the five stations might have a monotonous experience, but the students would learn important science process skills. (Mitchell, 1992, p. 62)

Mitchell is not advocating teaching to the manipulative section of the ESPET. Her point is that such instruction would not be fragmentary or isolated from a larger purpose, as would be the case if a learner were prepared for a specific multiple-choice or fill-in test. Important skills would be mastered, which could lead to further learning.

## Scoring the ESPET

The five stations in the manipulative skills section of the ESPET require a total of 18 responses. For the station requiring measurements of weight, volume, temperature, and height the test developers established a range of acceptable responses. Answers within this range received 1 point. All others are scored 0 with no partial credit allowed.

At the station that tests prediction, learners are expected to drop water on papers of varying absorbency and then predict what would happen on a paper they could not see or experiment with. Their predictions receive differential weighting: three points for describing (within a given range) what happened when the water was dropped, 1 point for predicting correctly, and 1 point for giving an acceptable reason. When scoring these responses, teachers must balance tendencies to generosity and justice, particularly when responses were vague or writing illegible.

The scoring standards are called **rubrics.** The classroom teachers are the raters. They are trained to compare a learner's answers with a range of acceptable answers prepared as guides. However, the rubrics acknowledge that these answers are *samples* of acceptable responses, rather than an exhaustive list. Thus, raters are required continually

to judge the quality of individual student answers. All ESPET scoring is done from student responses recorded on answer sheets.

## Protecting Scoring Reliability

Performance tests such as the ESPET, which require large numbers of raters across schools and classrooms, must be continually monitored to protect the reliability of the ratings. That is, the science achievement scores for learners in different fourth grades in different schools or school districts should be scored comparably. There are several ways to accomplish this.

For some performance tests, a representative sample of tests is rescored by the staff of another school. Sometimes teachers from different schools and school districts get together and score all their examinations in common. In other cases, a “recalibration” process is used, whereby individual graders pause in the middle of their grading and grade a few tests together as a group to ensure that their ratings are not drifting away from a common standard. We will describe this process in more detail in the next section.

## Community Accountability

Performance tests such as the ESPET do not have statewide or national norms that allow comparison with other learners in order to rank the quality of achievement. How, then, does a parent or school board know that the learning demonstrated on a science or math performance test represents a significant level of achievement? How does the community know that standards haven’t simply been lowered or that the learner is not being exposed to new but possibly irrelevant content?

The answer lies with how the content for a performance test is developed. For example, the syllabus on which the ESPET is based was developed under the guidance of experts in the field of science and science teaching. In addition, the recalibration

process ensures that science teachers at one school or school district will read and score examinations from other schools or school districts. Teachers and other professionals in the field of science or math can be expected to be critics for one another, ensuring that the syllabus will be challenging and the tests graded rigorously.

Experience with standardized performance testing in science and history in New York State and in mathematics and writing in California and Arizona has shown that cross-grading between schools and school districts provides some assurance that student learning as demonstrated on performance tests represents something of importance beyond the test-taking skills exhibited in the classroom. Nevertheless, as we will see next, research examining the cognitive complexity, validity, reliability, transfer, generalizability, and fairness of teacher-made, statewide, or national performance tests has only just begun (Herman, 1992).

## What Research Suggests About Performance Tests

Some educators (for example, Herman, 1992) believe that traditional multiple-choice exams have created an overemphasis on basic skills and a neglect of thinking skills in American classrooms. Now that several states have implemented performance tests, is there any evidence that such tests have increased the complexity of cognitive goals and objectives? Herman (1992) reports that California's eighth-grade writing assessment program, which includes performance tests based on portfolios, has encouraged teachers to require more and varied writing of their learners. In addition, the students' writing skills have improved over time since these new forms of assessment were first implemented. Mitchell (1992) reports that since the ESPET was begun in 1989, schools throughout New York State have revamped their science curricula to include thinking skills for all learners.

Both Herman and Mitchell, however, emphasize that the development of performance tests without parallel improvements in curricula can result in undesirable or inefficient

instructional practices, such as teachers drilling students on performance test formats. In such cases, improved test performance will not indicate improved thinking ability, nor will it generalize to other measures of achievement (Koretz, Linn, Dunbar, & Shepard, 1991).

## Do Performance Tests Measure Generalizable Thinking Skills?

Although tests such as the ESPET appear to assess complex thinking, research into their validity has just begun. While the recent developments in cognitive learning reviewed in Chapter 5 have influenced the developers of performance tests (Resnick & Klopfer, 1989; Resnick & Resnick, 1991), there is no conclusive evidence at present to suggest that important metacognitive and affective skills are being learned and generalized to tasks and situations that occur outside the performance test format and classroom (Linn, Baker, & Dunbar, 1991).

Shavelson and his colleagues (Shavelson, Gao, & Baxter, 1991) caution that conclusions drawn about a learner's problem-solving ability on one performance test may not hold for performance on another set of tasks. Similarly, Gearheat, Herman, Baker, and Wittaker (1992) have pointed out the difficulties in drawing conclusions about a learner's writing ability based on portfolios that include essays, biographies, persuasive writing, and poetry, which can indicate substantial variation in writing skill depending on the type of writing undertaken. We will return to the assessment of student portfolios later in the chapter.

## Can Performance Tests Be Scored Reliably?

Little research into the technical quality of standardized performance tests has been conducted. Nevertheless, current evidence on the ability of teachers and other raters to reliably and efficiently score performance tests is encouraging (Herman, 1992). Studies of portfolio ratings in Vermont, science scoring in New York, and hands-on math

assessment in Connecticut and California suggest that large-scale assessments can be administered and reliably scored by trained teachers working individually or in teams.

## Summary

Statewide performance tests have been developed, administered, and reliably scored for a number of years. National panels and study groups are developing a set of standardized performance exams that all American students will take in grades 4, 8, and 12 (Resnick & Resnick, 1991). It is not yet clear whether performance tests will become as common an assessment tool in American classrooms as traditional forms of assessment are now.

Nevertheless, many developments in the design of performance tests are occurring at a rapid pace. Curriculum and measurement experts have developed tests at the statewide level that can be reliably scored and efficiently administered by teachers. These tests have encouraged more complex learning and thinking skills and in some cases, as in New York, have led to performance-based revisions of the curriculum. Accounts by Mitchell (1992), Wiggins (1992), and Wolf, LeMahieu, and Eresh (1992) suggest that teachers who have used performance tests report improved thinking and problem solving in their learners. Also, school districts in Colorado, Oregon, California, New York, New Hampshire, Texas, Illinois, and other states have taken it on themselves to experiment with performance tests in their classrooms (*Educational Leadership*, 1992). In the next section we will present a process for developing, scoring, and grading performance tests based on the cumulative experience of these teachers and educators.

## Developing Performance Tests for Your Learners

Four years ago, Crow Island Elementary School began a project which has reaped benefits far beyond what any of us could have imagined. The focus of the project

> was assessment of children's learning, and the tangible product is a new reporting form augmented by student portfolios....The entire process has been a powerful learning experience....We are encouraged to go forward by the positive effects the project has had on the self-esteem and professionalism of the individual teachers and the inevitable strengthening of the professional atmosphere of the entire school. We have improved our ability to assess student learning. Equally important, we have become, together, a more empowered, effective faculty. (Hebert, 1992, p. 61)

> Brian doesn't like to write. Brian doesn't write. When Brian does write, it's under duress, and he doesn't share this writing. Last year I began working with a technique called portfolio assessment....Over the year Brian began to write and share his writing with others. His portfolio began to document success rather than failure. His voice, which has always been so forceful on the playground, had begun to come through in his writing as well. (Frazier & Paulson, 1992, p. 65)

As we learned in the previous section, performance assessment has the potential to improve both instruction and learning. As the quotations above illustrate, many educators around the country have decided to give it a try. What these educators have found is that performance assessment has not replaced traditional paper-and-pencil tests. Rather, it has supplemented these measures with tests that allow learners to demonstrate thinking skills through writing, speaking, projects, demonstrations, and other observable actions.

But as we have also learned, there are both conceptual and technical issues associated with the use of performance tests that teachers must resolve before performance assessments can be effectively and efficiently used. In this section we will discuss some of the important considerations in planning and designing a performance test. We will identify the tasks around which performance tests are based, and describe how to

develop a set of scoring rubrics for these tasks. Also included in this section will be suggestions on how to improve the reliability of performance test scoring, including portfolios. Figure 13.2 shows the major steps in building a performance test. We discuss each step in detail below.

## Deciding What to Test

Performance tests, like all authentic tests, are syllabus- or performance objectives-driven. Thus, the first step in developing a performance test is to create a list of objectives that specifies the knowledge, skills, attitudes, and indicators of these outcomes, which will then be the focus of your instruction. There are three general questions to ask when deciding what to test:

1. What knowledge or content (i.e., facts, concepts, principles, rules) is essential for learner understanding of the subject matter?
2. What intellectual skills are necessary for the learner to use this knowledge or content?
3. What habits of mind or attitudes are important for the learner to successfully perform with this knowledge or content?

Performance objectives that come from answering question 1 are usually measured by paper-and-pencil tests (discussed in Chapter 12). Objectives derived from answering questions 2 and 3, although often assessed with objective or essay-type questions, can be more authentically assessed with performance tests. Thus, your assessment plan for a unit should include both paper-and-pencil tests, to measure mastery of content, and performance tests, to assess skills and attitudes. Let’s see what objectives for these latter outcomes might look like.

### Performance Objectives in the Cognitive Domain.

Designers of performance tests usually ask these questions to help guide their initial selection of objectives:

- What kinds of essential tasks, achievements, or other valued competencies am I missing with paper-and-pencil tests?
- What accomplishments of those who practice my discipline (historians, writers, scientists, mathematicians) are valued but left unmeasured by conventional tests?

Typically, two categories of intellectual skills are identified from such questions: (a) skills related to acquiring information, and (b) skills related to organizing and using information. The accompanying box, *Designing a Performance Test,* contains a suggested list of skills for acquiring, organizing, and using information. As you study this list, consider which skills you might use as a basis for a performance test in your area of expertise. Then study the list of sample objectives in the bottom half of the box, and consider how these objectives are related to the list of skills.

Performance Objectives in the Affective and Social Domain. Performance assessments require the curriculum not only to teach thinking skills but also to develop positive dispositions and habits of mind. Habits of mind include such behaviors as constructive criticism, tolerance of ambiguity, respect for reason, and appreciation for the significance of the past. Performance tests are ideal vehicles for assessing positive attitudes toward learning, habits of mind, and social skills (for example, cooperation, sharing, and negotiation). Thus, in deciding what objectives to teach and measure with a performance test, you should give consideration to affective and social skill objectives. The following are some key questions to ask for including affective and social skills in your list of performance objectives:

- What dispositions, attitudes, or values characterize successful individuals in the community who work in your academic discipline?
- What are some of the qualities of mind or character traits possessed by good scientists, writers, reporters, historians, mathematicians, musicians, and so on?

- What will I accept as evidence that my learners have developed or are developing these qualities?
- What social skills for getting along with others are necessary for success as a journalist, weather forecaster, park ranger, historian, economist, mechanic, and so on?
- What evidence will convince my learners' parents that their children are developing these skills?

The accompanying box, *Identifying Attitudes for Performance Assessment,* displays some examples of attitudes, or habits of mind, that could be the focus of a performance assessment in science, social studies, and mathematics. Use it to select attitudes to include in your design of a performance assessment in these areas.

In this section, we illustrated the first step in designing a performance test by helping you identify the knowledge, skills, and attitudes that will be the focus of your instruction and assessment. The next step is to design the task or context in which these outcomes will be assessed.

## Designing the Assessment Context

The purpose of this step is to create an authentic task, simulation, or situation that will allow learners to demonstrate the knowledge, skills, and attitudes they have acquired. Ideas for these tasks may come from newspapers, reading popular books, or interviews with professionals as reported in the media (for example, an oil tanker runs aground and creates an environmental crisis, a drought occurs in an underdeveloped country causing famine, a technological breakthrough presents a moral dilemma). The tasks should center on issues, concepts, or problems that are relevant to your subject matter. In other words, they should be the same issues, concepts, and problems faced every day by important people working in the field.

Here are some questions to get you started, suggested by Wiggins (1992):

- What does the doing of mathematics, history, science, art, writing, and so forth look and feel like to professionals who make their living working in these fields in the real world?
- What are the projects and tasks performed by these professionals that can be adapted to school instruction?
- What roles—or habits of mind—do these professionals acquire that learners can re-create in the classroom?

The tasks you create may involve debates, mock trials, presentations to a city commission, reenactments of historical events, science experiments, or job responsibilities (for example, a travel agent, weather forecaster, or park ranger). Regardless of the specific context, they should present the learner with an authentic challenge. For example, consider the following social studies performance test (adapted from Wiggins, 1992):

> You and several travel agent colleagues have been assigned the responsibility of designing a trip to China for 12- to 14-year-olds. Prepare an extensive brochure for a month-long cultural exchange trip. Include itinerary, modes of transportation, costs, suggested budget, clothing, health considerations, areas of cultural sensitivity, language considerations, and other information necessary for parents to decide whether they want their child to participate.

Notice that this example presents learners with the following:

1. A hands-on exercise or problem to solve, which produces
2. an observable outcome or product (typed business letter, a map, graph, piece of clothing, multi-media presentation, poem, and so forth), which enables the teacher to

3. observe and assess not only the product but also the process used to arrive at it.

Designing the content for a performance test involves equal parts inspiration and perspiration. While no formula or recipe guarantees a valid performance test, the criteria given here can help guide you in revising and refining the task (Resnick & Resnick, 1991; Wiggins, 1992).

**1. Make the requirements for task mastery clear, but not the solution.** While your tasks should be complex, the required final product should be clear. Learners should not have to question whether they have finished or provided what the teacher wants. They should, however, have to think long and hard about how to complete the task. As you refine the task, make sure you can visualize what mastery of the task looks like and identify the skills that can be inferred from it.

**2. The task should represent a valid sample from which generalizations about the learner's knowledge, thinking ability, and attitudes can be made.** What performance tests lack in breadth of coverage they can make up in depth. In other words, they force you to observe a lot of behavior in a narrow domain of skill. Thus, the tasks you choose should be complex enough and rich enough in detail to allow you to draw conclusions about transfer and generalization to other tasks. Ideally, you should be able to identify 8 to 10 important performance tasks for an entire course of study (one or two per unit) that assess the essential performance outcomes you want your learners to achieve (Shavelson & Baxter, 1992).

**3. The tasks should be complex enough to allow for multimodal assessment.** Most assessment tends to depend on the written word. Performance tests, however, are designed to allow learners to demonstrate learning through a variety of modalities. This is referred to as **multimodal assessment.** In science, for example, one could make direct observations of students while they investigate a problem using laboratory equipment, give oral explanations of what they did, record procedures and conclusions in notebooks,

prepare exhibits of their projects, and solve short-answer paper-and-pencil problems. A multimodal assessment of this kind is more time-consuming than a multiple-choice test, but it will provide unique information about your learners' achievements untapped by other assessment methods. Shavelson and Baxter (1992) have shown that performance tests allow teachers to draw different conclusions about a learner's problem-solving ability than do higher-order multiple-choice tests or restricted-response essay tests, which ask learners to analyze, interpret, and evaluate information.

**4. The tasks should yield multiple solutions where possible, each with costs and benefits.** Performance testing is not a form of practice or drill. It should involve more than simple tasks for which there is one solution. Performance tests should, in the words of Resnick (1987), be nonalgorithmic (the path of action is not fully specified in advance), be complex (the total solution cannot be seen from any one vantage point), and involve judgment and interpretation.

**5. The tasks should require self-regulated learning.** Performance tests should require considerable mental effort and place high demands on the persistence and determination of the individual learner. The learner should be required to use cognitive strategies to arrive at a solution rather than depend on coaching at various points in the assessment process.

We close this section with three boxes, *Designing a Performance Assessment: Math, Communication,* and *History.* Each contains an example of a performance assessment task that contains most of these design considerations. Note that the first of these, the math assessment, also contains a scoring rubric, which is the subject of our next section.

## Specifying the Scoring Rubrics

One of the principal limitations of performance tests is the time required to score them reliably. Just as these tests require time and effort on the part of the learner, they demand a similar commitment from teachers when scoring them. True-false, multiple-

choice, and fill-in items are significantly easier to score than projects, portfolios, or performances. In addition, these latter accomplishments force teachers to make difficult choices about how much qualities like effort, participation, and attitude count in the final score.

Given the challenges confronting teachers who use performance tests, there is a temptation to limit the scoring criteria to the qualities of performance that are easiest to rate (e.g., keeping a journal of problems encountered) rather than the most important required for doing an effective job (e.g., the thoroughness with which the problems encountered were resolved). Wiggins (1992) cautions teachers that scoring the easiest or least controversial qualities can turn a well-thought-out and authentic performance test into a bogus one. Thus, your goal when scoring performance tests is to do justice to the time spent developing them and the effort expended by students taking them. You can accomplish this by developing carefully articulated scoring systems, or rubrics.

By giving careful consideration to rubrics, you can develop a scoring system for performance tests that minimizes the arbitrariness of your judgments while holding learners to high standards of achievement. Following are some of the important considerations in developing rubrics for a performance test.

Developing Rubrics. You should develop scoring rubrics that fit the kinds of accomplishments you want to measure. In general, performance tests require four types of accomplishments from learners:

- **Products:** Poems, essays, charts, graphs, exhibits, drawings, maps, and so forth.
- **Complex cognitive processes:** Skills in acquiring, organizing, and using information.
- **Observable performances:** Physical movements, as in dance, gymnastics, or typing; oral presentations; use of specialized equipment, as in focusing a

microscope; following a set of procedures, as when dissecting a frog, bisecting an angle, or following a recipe.

- **Attitudes and social skills:** Habits of mind, group work, and recognition skills.

As this list suggests, the effect of your teaching may be realized in a variety of ways. The difficulty in scoring some of these accomplishments should not deter your attempts to measure them. Shavelson and Baxter (1992), Kubiszyn and Borich (1996), and Sax (1989) have shown that if they are developed carefully and the training of those doing the scoring has been adequate, performance measures can be scored reliably.

Choosing a Scoring System. Choose a scoring system best suited for the type of accomplishment you want to measure. In general, there are three categories of rubrics to use when scoring performance tests: checklists, rating scales, and holistic scoring (see Figure 13.3). Each has certain strengths and limitations, and each is more or less suitable for scoring products, cognitive processes, performances, or attitudes and social skills.

**Checklists.** Checklists contain lists of behaviors, traits, or characteristics that can be scored as either present or absent. They are best suited for complex behaviors or performances that can be divided into a series of clearly defined specific actions. Dissecting a frog, bisecting an angle, balancing a scale, making an audiotape recording, or tying a shoe are behaviors requiring sequences of actions that can be clearly identified and listed on a checklist. Checklists are scored on a yes/no, present/absent, 0 or 1 point basis and should also allow the observer to indicate that she had no opportunity to observe the performance. Some checklists also list common mistakes that learners make when performing the task. In such cases, a score of +1 may be given for each positive behavior, –1 for each mistake, and 0 for no opportunity to observe. Figures 13.4 and 13.5 show checklists for using a microscope and a calculator.

**Rating Scales.** Rating scales are typically used for aspects of a complex performance that do not lend themselves to the yes/no or present/absent type of judgment. The most common form of rating scale is one that assigns numbers to categories of performance. Figure 13.6 (p. 449) shows a rating scale for judging elements of writing in a term paper. This scale focuses the rater's observations on certain aspects of the performance (accuracy, logic, organization, style, and so on) and assigns numbers to five degrees of performance.

Most numeric rating scales use an analytical scoring technique called **primary trait scoring** (Sax, 1989). This type of rating requires that the test developer first identify the most salient characteristics, or traits of greatest importance, when observing the product, process, or performance. Then, for each trait, the developer assigns numbers (usually 1–5) that represent degrees of performance.

Figure 13.7 (p. 450) displays a numerical rating scale that uses primary trait scoring to rate problem solving (Szetela & Nicol, 1992). In this system, problem solving is subdivided into the primary traits of understanding the problem, solving the problem, and answering the problem. For each trait, points are assigned to certain aspects or qualities of the trait. Notice how the designer of this rating scale identified characteristics of both effective and ineffective problem solving.

Two key questions are usually addressed in the design of scoring systems for rating scales using primary trait scoring (Wiggins, 1992):

- What are the most important characteristics that show a high degree of the trait?
- What are the errors most justifiable for achieving a lower score?

Answering these questions can prevent raters from assigning higher or lower scores on the basis of performance that may be trivial or unrelated to the purpose of the performance test, such as the quantity rather than the quality of a performance. One of the advantages of rating scales is that they focus the scorer on specific and relevant

aspects of a performance. Without the breakdown of important traits, successes, and relevant errors provided by these scales, a scorer's attention may be diverted to aspects of performance that are unrelated to the purpose of the performance test.

**Holistic Scoring. Holistic scoring** is used when the rater estimates the overall quality of the performance and assigns a numerical value to that quality, rather than assigning points for specific aspects of the performance. Holistic scoring is typically used in evaluating extended essays, term papers, or artistic performances, such as dance or musical creations. For example, a rater might decide to score an extended essay question or term paper on an A–F rating scale. In such a case, it would be important for the rater to have a model paper that exemplifies each score. After having created or selected these models from the set to be scored, the rater again reads each paper and then assigns each to one of the categories. A model paper for each category (A–F) helps to assure that all the papers assigned to a given category are of comparable quality.

Holistic scoring systems can be more difficult to use for performances than for products. For the former, some experience in rating the performance (for example, dramatic rendition, oral interpretations, or debate) may be required. In these cases, audiotapes or videotapes from past classes can be helpful as models representing different categories of performance.

**Combined Scoring Systems.** As was suggested, good performance tests require learners to demonstrate their achievements through a variety of primary traits, such as cooperation, research, and delivery. In some cases, therefore, the best way to arrive at a total assessment may be to combine several ratings, from checklists, rating scales, and holistic impressions. Figure 13.8 shows how scores across several traits for a current events project might be combined to provide a single performance score.

**Comparing the Three Scoring Systems.** Each of the three scoring systems has strengths and weaknesses. Table 13.2 serves as a guide in choosing a particular scoring system for a given type of performance, according to the following criteria:

- **Ease of construction:** the time involved in coming up with a comprehensive list of the important aspects or traits of successful and unsuccessful performance. Checklists, for example, are particularly time-consuming, while holistic scoring is not.
- **Scoring efficiency:** the amount of time required to score various aspects of the performance and calculate these scores as an overall score.
- **Reliability:** the likelihood that two raters will independently come up with a similar score, or the likelihood that the same rater will come up with a similar score on two separate occasions.
- **Defensibility:** the ease with which you can explain your score to a student or parent who challenges it.
- **Quality of feedback:** the amount of information the scoring system gives to learners or parents about the strengths and weaknesses of the performance.

Assigning Point Values. When assigning point values to various aspects of the performance test, it is a good idea to limit the number of points the assessment or component of the assessment is worth to that which can be reliably discriminated. For example, if you assign 25 points to a particular product or procedure, you should be able to distinguish 25 degrees of quality. When faced with more degrees of quality than can be detected, a typical rater may assign some points arbitrarily, reducing the reliability of the assessment.

On what basis should points be assigned to a response on a performance test? On the one hand, you want a response to be worth enough points to allow you to differentiate

subtle differences in response quality. On the other hand, you want to avoid assigning too many points to a response that does not lend itself to complex discriminations. Thus, assigning one or two points to a math question requiring complex problem solving would not allow you to differentiate between outstanding, above average, average, and poor responses. But assigning 30 points to this same answer would seriously challenge your ability to distinguish a rating of 15 from a rating of 18. Two considerations can help in making decisions about the size and complexity of a rating scale.

First, the scoring model should allow that rater to specify the exact performance—or examples of acceptable performance—that correspond with each scale point. The ability to successfully define distinct criteria, then, can determine the number of scale points that are defensible. Second, although it is customary for homework, paper-and-pencil tests, and report cards to use a 100 point (percent) scale, scale points derived from performance assessments do not need to add up to 100. We will have more to say later about assigning marks to performance tests and how to integrate them with other aspects of an overall grading system (for example, homework, paper-and-pencil tests, classwork), including portfolios.

## Specify Testing Constraints

Should performance tests have time limits? Should learners be allowed to correct their mistakes? Can they consult references or ask for help from other learners? Performance tests confront the designer with the following dilemma: If the test is designed to confront learners with real-world challenges, why shouldn’t they be allowed to tackle these challenges as real-world people do? In the world outside the classroom, mathematicians make mistakes and correct them, journalists write first drafts and revise them, weather forecasters make predictions and change them. Each of these workers can consult references and talk with colleagues. Why, then, shouldn’t learners who are

working on performance tests that simulate similar problems be allowed the same working (or testing) conditions?

Even outside the classroom, professionals have constraints on performance, such as deadlines, limited office space, or outmoded equipment. So how does a teacher decide which conditions to impose during a performance test? Before examining this question, let's look at some of the typical conditions, or **testing constraints,** imposed on learners during tests. Wiggins (1992) includes the following among the most common forms of testing constraints:

- **Time.** How much time should a learner have to prepare, rethink, revise, and finish a test?
- **Reference material.** Should learners be able to consult dictionaries, textbooks, or notes as they take a test?
- **Other people.** May learners ask for help from peers, teachers, and experts as they take a test or complete a project?
- **Equipment.** May learners use computers or calculators to help them solve problems?
- **Prior knowledge of the task.** How much information about the test situation should learners receive in advance?
- **Scoring criteria.** Should learners know the standards by which the teacher will score the assessment?

Wiggins recommends that teachers take an "authenticity test" to decide which of these constraints to impose on a performance assessment. His authenticity test involves answering the following questions:

- What kinds of constraints authentically replicate the constraints and opportunities facing the performer in the real world?

- What kinds of constraints tend to bring out the best in apprentice performers and producers?
- What are the appropriate or authentic limits one should impose on the availability of the six resources just listed?

Indirect forms of assessment, by the nature of the questions asked, require numerous constraints during the testing conditions. Allowing learners to consult reference materials or ask peers for help during multiple-choice tests would significantly reduce their validity. Performance tests, on the other hand, are direct forms of assessment in which real-world conditions and constraints play an important role in demonstrating the competencies desired.

## Portfolio Assessment

According to Paulson and Paulson (1991), portfolios "tell a story." The story, viewed as a whole, answers the question, "What have I learned during this period of instruction and how have I put it into practice?" Thus **portfolio assessment** is assessment of a learner's entire body of work in a defined area, such as writing, science, or math. The object of portfolio assessment is to demonstrate the student's growth and achievement.

Some portfolios represent the student's own selection of products—scripts, musical scores, sculpture, videotapes, research reports, narratives, models, and photographs—that represent the learner's attempt to construct his or her own meaning out of what has been taught. Other portfolios are preorganized by the teacher to include the results of specific products and projects, the exact nature of which may be determined by the student.

Whether portfolio entries are preorganized by the teacher or left to the discretion of the learner, several questions must be answered prior to the portfolio assignment:

- What are the criteria for selecting the samples that go into the portfolio?

- Will individual pieces of work be evaluated as they go into the portfolio, or will all the entries be evaluated collectively at the end of a period of time—or both?
- Will the amount of student growth, progress, or improvement over time be graded?
- How will different entries, such as videos, essays, artwork, and reports, be compared and weighted?
- What role will peers, parents, other teachers, and the student him- or herself have in the evaluation of the portfolio?

Shavelson, Gao, and Baxter (1991) suggest that at least eight products or tasks over different topic areas may be needed to obtain a reliable estimate of performance from portfolios. Therefore, portfolios are usually built and assessed cumulatively over a period of time. These assessments determine the quality of individual contributions to the larger portfolio at various time intervals and the quality of the entire portfolio at the end of instruction.

Various schemes have been devised for evaluating portfolios (Paulson & Paulson, 1991). Most involve a recording form in which (1) the specific entries are cumulatively rated over a course of instruction, (2) the criteria with which each entry is to be evaluated are identified beforehand, and (3) an overall rating scale is provided for rating each entry against the criteria given.

Frazier and Paulson (1992) and Hebert (1992) report successful ways in which peers, parents, and students themselves have participated in portfolio evaluations. Figure 13.9 represents one example of a portfolio assessment form intended for use as a cumulative record of accomplishment over an extended course of study.

Many teachers use portfolios to increase student reflections about their own work and encourage the continuous refinement of portfolio entries. Portfolios have the advantage of containing multiple samples of student work completed over time that can represent

finished works as well as works in progress. Entries designated “works in progress” are cumulatively assessed at regular intervals on the basis of student growth or improvement and on the extent to which the entry increasingly matches the criteria given.

## Performance Tests and Report Card Grades

Performance tests require a substantial commitment of teacher time and learner-engaged time. Consequently, the performance test grade should have substantial weight in the report card grade. Here are two approaches to designing a grading system that includes performance assessments.

One approach to scoring quizzes, tests, homework assignments, performance assessments, and so forth. is to score each on the basis of 100 points. Computing the final grade, then, simply involves averaging the grades for each component, multiplying these averages by the weight assigned, and adding these products to determine the total grade. The box titled *Using Grading Formulas* at the end of Chapter 12 provided examples of three formulas for accomplishing this. But as discussed above, these methods require that you assign the same number of points (usually 100) to everything you grade.

Another approach is to use a “percentage of total points” system. With this system you decide how many points each component of your grading system is worth on a case-by-case basis. For example, you may want some tests to be worth 40 points, some 75, depending on the complexity of the questions and the performance desired. Likewise, some of your homework assignments may be worth only 10 or 15 points. The accompanying box, *Using a Combined Grading System,* describes procedures involved in setting up such a grading scheme for a six-week grading period. Table 13.3 and Figure 13.10 provide some example data for how such a system works.

## Final Comments

Performance assessments create challenges that restricted-response tests do not. Performance grading requires greater use of judgment than do true-false or multiple-choice questions. These judgments can become more reliable if (1) the performance to be judged is clearly defined, (2) the ratings or criteria used to make the judgments are determined beforehand, and (3) two or more raters independently grade the performance and an average is taken.

Using videotapes or audiotapes can enhance the validity of performance assessments when direct observation of performance is required. Furthermore, performance assessments need not take place at one time for the whole class. Learners can be assessed at different times, individually or in small groups. For example, learners can rotate through classroom learning centers (Shalaway, 1989) and be assessed when the teacher feels they are acquiring mastery.

Finally, don't lose sight of the fact that performance assessments are meant to serve and enhance instruction rather than being simply an after-the-fact test given to assign a grade. When tests serve instruction, they can be given at a variety of times and in as many settings and contexts as instruction requires. Some performance assessments can sample the behavior of learners as they receive instruction or be placed within ongoing classroom activities rather than consume extra time during the day.

## Summing Up

This chapter introduced you to performance-based assessment. Its main points were these:

- Performance tests use direct measures of learning that require learners to analyze, problem solve, experiment, make decisions, measure, cooperate with others, present orally, or produce a product.
- Performance tests can assess not only higher-level cognitive skills but also noncognitive outcomes, such as self-direction, ability to work with others, and social awareness.
- Rubrics are scoring standards composed of model answers that are used to score performance tests. They are samples of acceptable responses against which the rater compares a student's performance.
- Research on the effects of performance assessment indicates that when teachers include more thinking skills in their lesson plans, higher levels of student performance tend to result. However, there is no evidence yet that the thinking skills measured by performance tests generalize to tasks and situations outside the performance test format and classroom.
- The four steps to constructing a performance assessment are deciding what to test, designing the assessment context, specifying the scoring rubrics, and specifying the testing constraints.
- A performance test can require four types of accomplishments from learners: products, complex cognitive processes, observable performance, and attitudes and social skills. These performances can be scored with checklists, rating scales, or holistic scales.
- Constraints that must be decided on when a performance test is constructed and administered are the amount of time allowed, use of reference material, help from others, use of specialized equipment, prior knowledge of the task, and scoring criteria.

- Two approaches to combining performance grades with other grades are (1) to assign 100 total points to each assignment that is graded and average the results, and (2) to use the percentage-of- total-point systems.

For Discussion and Practice

*1. Compare and contrast some of the reasons for giving conventional tests with those for giving performance assessments.

*2. Using an example from your teaching area, explain the difference between direct and indirect measures of behavior.

3. Describe some habits of mind that might be required by a performance test in your teaching area. How did you learn about the importance of these attitudes, social skills, and ways of working?

*4. Describe how at least two school districts have implemented performance assessments. Indicate the behaviors they assess and by what means they are measured.

5. Would you agree or disagree with this statement: “An ideal performance test is a good teaching activity”? With a specific example in your teaching area, illustrate why you answered as you did.

6. List at least two learning outcomes and describe how you would measure them in your classroom to indicate that a learner is (1) self-directed, (2) a collaborative worker, (3) a complex thinker, (4) a quality producer, and (5) a community contributor.

*7. Describe what is meant by a scoring rubric and how such rubrics were used in New York State’s Elementary Science Performance Evaluation Test.

*8. What two methods have been used successfully to protect the scoring reliability of a performance test? Which would be more practical in your own teaching area or at your grade level?

*9. What is meant by the community accountability of a performance test and how can it be determined?

*10. In your own words, how would you answer a critic of performance tests who says they don't measure generalizable thinking skills outside the classroom and can't be scored reliably?

11. Identify for a unit you will be teaching several attitudes, habits of mind, and/or social skills that would be important to using the content taught in the real world.

12. Create a performance test of your own choosing that (1) requires a hands-on problem to solve and (2) results in an observable outcome for which (3) the process used by learners to achieve the outcome can be observed. Use the five criteria by Wiggins (1992) and Resnick and Resnick (1991) to help guide you.

13. For the performance assessment above, describe and give an example of the accomplishments—or rubrics—you would look for in scoring the assessment.

14. For this same assessment, compose a checklist, rating scale, or holistic scoring method by which a learner's performance would be evaluated. Explain why you chose that scoring system, which may include a combination of the above methods.

15. For your performance assessment above, describe the constraints you would place on your learners pertaining to the time to prepare for and complete the activity; references that may be used; people that may be consulted, including other students; equipment allowed; prior knowledge about what is expected;

and points or percentages you would assign to various degrees of their performance.

16. Imagine you have to arrive at a final grade composed of homework, objective tests, performance tests, portfolio, classwork, and notebook, which together you want to add up to 100 points. Using Figure 13.10 and Table 13.3 as guides, compose a grading scheme that indicates the weight, number, individual points, and total points assigned to each component. Indicate the percentage of points required for the grades A–F.

Suggested Readings

ASCD (1992). Using performance assessment [special issue]. *Educational Leadership, 49*(8). This special issue contains clear, detailed examples of what teachers around the country are doing to give performance tests a try.

Linn, R. L., Baker, E., & Dunbar, S. B. (1991). Complex performance based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21. A clear, concise review of the strengths and limitations of performance tests. Also discusses the research that needs to be done to improve their validity and reliability.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: Free Press. The first comprehensive treatment of alternative approaches to traditional testing. Includes excellent discussions of the problems of current testing practice and the advantages of performance tests. The examples of performance tests are especially helpful.

**Performance testing.** Tests that use direct measures of learning rather than indicators that suggest that learning has taken place.

Can complex cognitive outcomes, such as critical thinking and decision making, be more effectively learned with performance tests than with traditional methods of testing?

A performance test can evaluate process as well as product and thus can measure more of what actually goes on in the classroom and in everyday life than can a pencil-and-paper test.

A good performance test is very much like a good lesson. During the test, learners have the opportunity to “show off” what they have been working hard to master.

**Figure 13.1**
Example of a performance activity and assessment.

How can I construct performance tests that measure self-direction, ability to work with others, and social awareness?

Table 13.1

Learning Outcomes of the Aurora Public Schools

**A Self-Directed Learner**

1. Sets priorities and achievable goals
2. Monitors and evaluates progress
3. Creates options for self

4. Assumes responsibility for actions
5. Creates a positive vision for self and future

**A Collaborative Worker**

6. Monitors own behavior as a group member
7. Assesses and manages group functioning
8. Demonstrates interactive communication
9. Demonstrates consideration for individual differences

**A Complex Thinker**

10. Uses a wide variety of strategies for managing complex issues
11. Selects strategies appropriate to the resolution of complex issues and applies the strategies with accuracy and thoroughness
12. Accesses and uses topic-relevant knowledge

**A Quality Producer**

13. Creates products that achieve their purpose
14. Creates products appropriate to the intended audience
15. Creates products that reflect craftsmanship
16. Uses appropriate resources/technology

**A Community Contributor**

17. Demonstrates knowledge about his or her diverse communities
18. Takes action
19. Reflects on his or her role as a community contributor

**Authentic assessment.** Testing that covers the content that was taught in the manner in which it was taught and that targets specific behaviors that have applicability to advanced courses, other programs of study, or careers.

Performance tests are worth studying for and teaching to. Well-designed performance tests can produce high levels of motivation and learning.

**Rubrics.** Scoring standards composed of model answers that are used to score performance tests.

Can standardized performance tests be scored reliably?

Performance tests challenge learners with real-world problems that require higher level cognitive skills to solve.

**Figure 13.2**
Steps for developing a performance test.

How do I decide what a performance test should measure?

Performance tests can be used to assess habits of mind, such as cooperation and social skills.

Applying Your Knowledge:

Designing a Performance Test

Following are lists of skills appropriate for performance tests in the cognitive domain. Below these lists are a number of sample performance-test objectives derived from the listed skills.

Skills in Organizing

Skills in Acquiring Information and Using Information

**Communicating**

explaining
modeling
demonstrating
graphing
displaying
writing
advising
programming
proposing
drawing

**Measuring**

counting
calibrating
rationing
appraising
weighing
balancing
guessing
estimating
forecasting

**Investigating**

gathering references
interviewing
using references

**Organizing**

classifying
categorizing
sorting
ordering
ranking
arranging

**Problem solving**

stating questions
identifying problems
developing hypotheses
interpreting
assessing risks
monitoring

**Decision making**

weighing alternatives
evaluating
choosing
supporting
defending
electing
adopting

experimenting

hypothesizing

- Write a summary of a current controversy drawn from school life and tell how a courageous and civic-minded American you have studied might decide to act on the issue.
- Draw a physical map of North America from memory and locate 10 cities.
- Prepare an exhibit showing how your community responds to an important social problem of your choosing.
- Construct an electrical circuit using wires, a switch, a bulb, resistors, and a battery.
- Describe two alternative ways to solve a mathematics word problem.
- Identify the important variables that accounted for recent events in our state, and forecast the direction they might take.
- Design a freestanding structure in which the size of one leg of a triangular structure must be determined from the other two sides.
- Program a calculator to solve an equation with one unknown.
- Design an exhibit showing the best ways to clean up an oil spill.
- Prepare a presentation to the city council, using visuals, requesting increased funding to deal with a problem in our community.

How do I design a performance test based on real-life problems important to people who are working in the field?

Applying Your Knowledge:

## Identifying Attitudes for Performance Assessment

Science*

- Desiring knowledge. Viewing science as a way of knowing and understanding.
- Being skeptical. Recognizing the appropriate time and place to question authoritarian statements and "self-evident truths."
- Relying on data. Explaining natural occurrences by collecting and ordering information, testing ideas, and respecting the facts that are revealed.
- Accepting ambiguity. Recognizing that data are rarely clear and compelling, and appreciating the new questions and problems that arise.
- Willingness to modify explanations. Seeing new possibilities in the data.
- Cooperating in answering questions and solving problems. Working together to pool ideas, explanations, and solutions.
- Respecting reason. Valuing patterns of thought that lead from data to conclusions and eventually to the construction of theories.
- Being honest. Viewing information objectively, without bias.

Social Studies†

- Understanding the significance of the past to their own lives, both private and public, and to their society.
- Distinguishing between the important and inconsequential to develop the "discriminating memory" needed for a discerning judgment in public and personal life.
- Preparing to live with uncertainties and exasperating, even perilous, unfinished business, realizing that not all problems have solutions.

- Appreciating the often tentative nature of judgments about the past, and thereby avoiding the temptation to seize on particular “lessons” of history as cures for present ills.

Mathematics‡

- Appreciating that mathematics is a discipline that helps solve real-world problems.
- Seeing mathematics as a tool or servant rather than something mysterious or mystical to be afraid of.
- Recognizing that there is more than one way to solve a problem.

*From Loucks-Horsley et al., 1990, p. 41.

†From Parker, 1991, p. 74.

‡From Willoughby, 1990.

**Multimodal assessment.** The evaluation of performance through a variety of forms.

Applying Your Knowledge:

Designing a Performance Assessment: Math

Joe, Sarah, José, Zabi, and Kim decided to hold their own Olympics after watching the Olympics on TV. They needed to choose the events to have at their Olympics. Joe and José wanted weight lifting and Frisbee toss events. Sarah, Zabi, and Kim thought a running event would be fun. The children decided to have all three events. They also decided to make each event of the same importance.

One day after school they held their Olympics. The children's mothers were the judges. The mothers kept the children's scores on each of the events.

The children's scores for each of the events are listed below:

| Child's Name | Frisbee Toss | Weight Lift | 50-Yard Dash |
|---|---|---|---|
| Joe | 40 yards | 205 pounds | 9.5 seconds |
| José | 30 yards | 170 pounds | 8.0 seconds |
| Kim | 45 yards | 130 pounds | 9.0 seconds |
| Sarah | 28 yards | 120 pounds | 7.6 seconds |
| Zabi | 48 yards | 140 pounds | 8.3 seconds |

Now answer the question "Who won the Olympics?" and give an explanation of how you arrived at your answer (4 points).

Sample Responses

**Student A**

Who would be the all-around winner?

ZABI

Explain how you decided who would be the all-around winner. Be sure to show all your work.

I decided by how each person came in and that is who won.

**Student B**

Who would be the all-around winner?

ZABI

Explain how you decided who would be the all-around winner. Be sure to show all your work.

I wrote in order all the scores from first place to fifth place. Then I added them up.

Whoever had the least amount won.

**Student C**

Who would be the all-around winner?

ZABI

Explain how you decided who would be the all-around winner. Be sure to show all your work.

Zabi got one first place and two third places. I counted 3 points for every first place they got and 2 points for second place and 1 point for third place. Zabi got the most points.

*Source:* From Blumberg, Epstein, MacDonald, & Mullis, 1986.

Applying Your Knowledge:

Designing a Performance Assessment: Communication

1. You are representing an ad agency. Your job is to find a client in the school who needs photos to promote his or her program. (Examples: the future teachers club, the fine arts program, Student Council.)
2. Your job is to research all the possibilities, select a program, learn about that program, and then record on film the excitement and unique characteristics that make up the program you have selected. Your photos will be used to advertise and stimulate interest in that area.
3. Visualize how you will illustrate your ideas either by writing descriptions or by drawing six of your proposed frames. Present these six ideas to your instructor (the director of the ad agency) before you shoot.

*Source:* Redding, 1992, p. 49.

Applying Your Knowledge:

Designing a Performance Assessment: History

You and your colleagues (groups of three or four) have been asked to submit a proposal to write a U.S. history textbook for middle school students. The publishers demand two things: that the book hit the most important events, and that it be interesting to students. Because of your expertise in eighteenth-century American history, you will provide them with a draft chapter on the eighteenth century, up to but not including the American Revolution, field-tested on some middle school students. They also ask that you fill in an "importance" chart with your responses to these questions:

1. Which event, person, or idea is most important in this time period, and why?
2. Which of three sources of history—ideas, people, events—is most important?

You will be expected to justify your choices of "most important" and to demonstrate that the target population is likely to be interested in your book.

*Source:* Wiggins, 1992, p. 28.

**Figure 13.3**

Types of scoring rubrics.

How can I use a simple checklist to score a performance test accurately and reliably?

What are some ways of using rating scales to score a performance test?

**Primary trait scoring.** An analytical scoring technique that requires a test developer to first identify the most salient characteristics or primary traits when observing a product, process, or performance.

**Figure 13.4**

Checklist for using a microscope.

**Figure 13.5**

Checklist for using an electronic calculator.

**Figure 13.6**

Rating scale for themes and term papers that emphasizes interpretation and organization.

**Figure 13.7**

Analytic scale for problem solving. *Source:* From Szetela & Nicol, 1992, p. 42.

**Holistic scoring.** Estimating the overall quality of a performance by giving a single value that represents a specific category of accomplishment.

**Figure 13.8**

Combined scoring rubric for a current events project: Total points equal 17.

Table 13.2

Comparison of Three Performance-Based Scoring Systems

| | Ease of Construction | Scoring Efficiency | Reliability | Defensibility | Feedback | More Suitable For |
|---|---|---|---|---|---|---|
| **Checklist** | low | moderate | high | high | high | procedures, attitudes |
| **Rating scale** | moderate | moderate | moderate | moderate | moderate | products, social skills |
| **Holistic scoring** | high | high | low | low | low | products and processes |

How do I decide how many total points to assign to a performance test?

How do I decide what conditions to place on my learners when completing a performance test to make it as authentic as possible?

**Testing constraints.** The amount of time, reference material, degree of help (if any) from others, specialized equipment, prior knowledge of the task, and scoring criteria that test-takers can have during a performance assessment.

**Portfolio assessment.** Assessment of a learner's entire body of work in a defined content area in order to demonstrate the student's growth and achievement.

What are student portfolios and how can they be graded fairly and objectively?

**Figure 13.9**
Example of portfolio assessment form. *Source:* Adapted from *Looking Beyond "The Answer," the Vermont Mathematics Portfolio, 1992*, report of the Vermont Mathematics Portfolio Assessment Program, publication year 1990–91. Montpelier, VT: Vermont Department of Education.

How do I weight performance tests and combine them with other student work, such as quizzes, homework, and class participation, to create a final grade?

Table 13.3

Example of Grade
Components and Weights

| **Component** | **Weight** |
|---|---|
| Homework | 15% |
| Objective tests | 20% |
| Performance tests | 20% |
| Portfolio | 20% |
| Classwork | 15% |
| Notebook | 10% |
| TOTAL | 100% |

**Figure 13.10**
Sample grade recording sheet, first marking period.

Applying Your Knowledge:

Using a Combined Grading System

**Step 1:** Identify the components of your grading system and assign each component a weight. A weight is the percentage of total points a particular component carries. Table 13.3 displays components and weights for a six-week grading plan.

**Step 2:** Record the actual points each student earned out of the number of points possible in the grade book. Leave a column for totals. (See Figure 13.10.) As you can see, each component and each separate assignment has varying numbers of

possible points that can be earned. Assign points to each component based on the complexity of the required performance, the length of the assignment, and your perception of your ability to assign reliable ratings.

**Step 3:** Total the actual points earned for each component and divide the total by the possible points. The results represent the percentage of points earned for each particular component. Thus, in our example from Figure 13.10, Cornell and Rosie earned the following points and totals:

| | **Cornell** | **Rosie** |
|---|---|---|
| Homework | 50/70=71% | 55/70=79% |
| Objective tests | 45/60=75% | 35/60=58% |
| Performance tests | 33/40=83% | 39/40=98% |
| Portfolio | 18/20=90% | 15/20=75% |
| Classwork | 39/50=78% | 37/50=74% |
| Notebook | 5/10=50% | 8/10=80% |

**Step 4:** Multiply each of these percentages by the weights assigned, as shown in Table 13.3, and total these products.

| | **Cornell** | **Rosie** |
|---|---|---|
| Homework | 713.15=10.6 | 793.15=11.8 |
| Objective tests | 753.20=15 | 583.20=11.6 |
| Performance tests | 833.20=16.6 | 983.20=19.6 |
| Portfolio | 903.20=18 | 753.20=15 |
| Classwork | 783.15=11.7 | 743.15=11.1 |
| Notebook | 503.10=5 | 803.10=8 |
| Totals | 76.9 | 77.1 |

**Step 5:** Record the grade either as a letter grade (A=90–100 percent, etc.) or as the percentage itself, depending on your school's system.

Questions marked with an asterisk are answered in the appendix.