CHAPTER

11

MEASURING COMPLEX ACHIEVEMENT: PERFORMANCE-BASED ASSESSMENTS

Objectives

From this chapter you should be able to:

1. Describe the uses of performance-based assessments.
2. Describe the advantages and limitations of performance-based assessments.
3. Distinguish between restricted-response and extended-response performance-based assessments.
4. Construct performance-based assessments.
5. Construct scoring rubrics, rating scales, and checklists for performance-based assessments.

Essay tests are the most common example of a performance-based assessment, but there are many others, including artistic productions, experiments in science, oral presentations, and the use of mathematics to solve real-world problems. The emphasis is on doing, not merely knowing—on process as well as product.

Essay tests are an example of one type of performance assessment. But there are many aspects of writing that are not tapped within the constraints of the normal essay test. Choosing a topic, identifying an audience, gathering information, preparing drafts, seeking critiques, collaborating with other students and revising are all important aspects of writing that are not measured by the usual essay test. Moreover, writing is not the only type of performance outcome we need to assess. Many highly valued learning outcomes emphasize the actual performance of tasks in realistic settings. This is obvious in the case of art or music and for vocational or industrial education courses, such as auto repair, woodworking, or typing. It is also true for mathematics, science, social studies, and

foreign languages. In each case, performance-based assessments are needed to measure some of the desired learning outcomes.

For example, while knowledge of vocabulary and grammar in a foreign language can be measured with the various forms of paper-and-pencil tests, speaking skills cannot. Oral performance is required to assess a student's spoken communication skills in a foreign language. Similarly, the assessment of a student's ability to make observations, formulate hypotheses, collect data, and draw valid scientific conclusions may require the use of performance assessments. The use of mathematics to solve meaningful real-world problems and to communicate solutions to others may also be best assessed by the use of performance tasks in realistic settings.

Performance assessments provide a basis for teachers to evaluate both the effectiveness of the *process* or procedure used (e.g., approach to data collection or manipulation of instruments) and the *product* resulting from performance of a task (e.g., completed report of results or completed artwork). Unlike simple tests of factual knowledge, there is unlikely to be a single right or best answer. Rather, there may be multiple performances and problem solutions that would be judged to be excellent. Problem formulation, the organization of ideas, the integration of multiple types of evidence, and originality are all important aspects of performance that may not be adequately assessed by paper-and-pencil tests.

TYPES OF PERFORMANCE-BASED ASSESSMENT

Performance assessments are also sometimes referred to as "authentic assessments" or "alternative assessments." But the terms are not interchangeable. "Alternative assessment" highlights the contrast to traditional paper-and-pencil tests, while "authentic assessment" emphasizes the practical application of the tasks in real-world settings. We prefer the label "performance assessment" because it is more descriptive than "alternative assessment" and less pretentious than "authentic assessment."

Authenticity is, obviously, a matter of degree. A highly authentic assessment of communication skills in German, for example, might involve listening to the verbal interactions of a student when visiting Germany. But such an assessment obviously would lack practicality for the teacher of a typical German class. Simulated spoken interactions between the teacher and a student or among students, while not quite as authentic, are much more practical. In either case, the focus of the assessment is on the student's performance in communicating in German.

Although authenticity is usually only approximated, it is an important goal of performance assessment. Providing realistic contexts can make problems more engaging for students and help the teacher evaluate whether a student who can solve a problem in one context can solve it in another. Hence, it is desirable to increase the authenticity of tasks to whatever extent possible.

Like essay questions, performance assessments should be used primarily to measure those learning outcomes that cannot be measured well by objective test items. Objective test items are generally more efficient and more reliable for measuring factual knowledge and the ability to solve well-structured problems (e.g., solve a quadratic equation). Performance assessments are better suited for applications with less structured problems where problem identification; collection, organization, integration, and evaluation of information; and originality are emphasized (e.g., where is the best place to locate a restaurant?).

They are also essential for learning outcomes that involve the creation of a product (e.g., painting) or an oral or physical performance (e.g., the presentation of a speech, the repair of an engine, or the use of a scientific instrument).

"Hands-on" performance tasks that require students to manipulate objects, measure outcomes, and observe results of experimental manipulations are sometimes essential to capture the full array of skills needed to perform "authentic" tasks. This is obvious in the case of a driving test or a performance test for a dentist, but it may also be true in science and other areas. Research has shown that computer simulations of tasks in science sometimes may be good substitutes for actual hands-on performance of the task, but in other instances even high-fidelity simulations may have relatively poor relationships for hands-on performance. Poor relationships between simulations and actual hands-on performance occur most commonly when the manipulation of apparatus (e.g., mixing a compound or taking a measurement) is an integral part of the task.

Performance tasks can vary substantially in the degree to which performance is restricted. Writing a program, for example, might be completely constrained with regard to the computer language and the purpose of the programming. The task of creating a sculpture might be almost completely unconstrained with regard to the approach a student might take or the nature of the product produced. Most performance tasks fall in between these extremes.

Restricted-Response Performance Tasks

A restricted-response performance task is usually relatively narrow in definition. The instructions are generally more focused than extended-response performance tasks, and the limitations on the types of performance expected are likely to be indicated.

Restricted-response performance tasks sometimes start with a simple multiple-choice or short-answer question, such as the one in Figure 11.1. Those questions are then extended by asking for an explanation of the answer and sometimes an explanation for why the other answers were not selected. Often, different answers in the first part of the task could be given full credit if the explanation provided sound reasoning to defend the choice.

EXAMPLES

Read aloud a section of a story.

Use various combinations of five straight pieces of plastic to construct as many different triangles as you can and record the perimeters of each.

Determine which of two liquids contains sugar and explain what results support your conclusion.

Construct graphs of the average amount of rainfall per month for two cities.

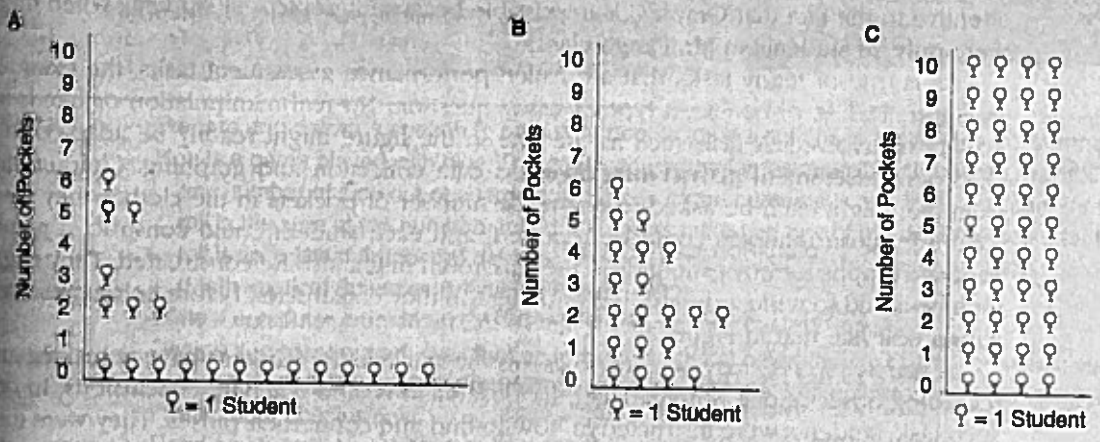
Request aloud directions to the train station in French.

Write the names of the countries in the appropriate areas of a blank map of Europe.

Sara knows that half the students in her class were invited to Kim's birthday party. Also, half were invited to Julie's party. Sara thinks that these figures add up to 100%, so she thinks she will surely be invited to one of the parties. Explain why Sara is wrong. If possible use a diagram in your explanation.*

*Adapted from a task used in the California Assessment program.

There are 20 students in Mr. Pang's class. On Tuesday most of the students in the class said they had pockets in the clothes they were wearing.



Which of the graphs most likely shows the number of pockets that each child had? ____

Explain why you chose that graph.

Explain why you didn't choose the other two graphs.

Figure 11.1
Example of stimulus material for a mathematics problem administered at grade 4 in the 1992 National Assessment of Educational Progress

Source: NAEP 1992: Mathematics Report Card for the Nation and the States (p. 49) by I. V. S. Mullis, J. A. Dossey, E. H. Owen, and G. W. Phillips, 1993. Washington, DC: U.S. Department of Education. Report No. 23-ST02.

If the explanation parts of the task in Figure 11.1 were omitted, there would be no way to determine the basis for a student's choice of one of the three figures. Even if students selected the preferred choice (B), you would not know whether they did so for a sound reason or whether they simply guessed. Nor would you know whether they were attentive to the fact that Graph C is impossible because it depicts 44 students when there were only 20 students in Mr. Pang's class.

As is true of many tasks that are called performance assessment tasks, the example in Figure 11.1 is, of course, a type of essay question. No real manipulation or hands-on activity is involved. A task such as the one in the figure might readily be adapted to a classroom assessment activity that involved data collection and graphing. Children, for example, might each be asked to count the number of pockets in the clothes they were wearing. Those numbers could be reported, and each student could construct a graph. Separate graphs for boys and girls in the classroom might also be constructed. They might then be asked to write a description of the graph they constructed before being presented with a task like that in Figure 11.1.

A variety of tasks may be used to assess the skills young students have at making and recording observations, summarizing the observations, and reaching conclusions. In one such task, students were instructed in how to find and count their pulses. They were then asked to count the number of pulses in each of four segments of 15 seconds where the teacher looked at a stopwatch and gave instructions to start and stop. After recording the four initial segments, students were told to jump up and down for 1 minute. After exercise, children were asked to count and record their pulses for four additional 15-second periods. Next, a second period of jumping for 2 minutes was required, followed by four more recordings of 15-second segments. Students were asked to construct a table and a graph reporting the results and describe what the results showed when the initial four recordings were compared to those following the first and second rounds of exercise. Finally, they were asked to explain what they observed.

The relative advantages and disadvantages of restricted performance tasks parallel those of restricted essay questions. They are generally more structured and require less time to administer than extended-response performance tasks. The shorter administration time makes it possible to administer more tasks and thereby gain broader coverage of the content domain. The greater degree of structure makes the task easier to score. On the other hand, the structure makes the tasks less valuable for measuring student skills such as approaches to ill-structured problems, integration of information, and originality. Extended performance tasks are better suited for such outcomes.

Extended Performance Tasks

The extended performance task may require students to seek information from a variety of sources beyond those provided by the task itself. For example, students may need to use the library, make observations, collect and analyze data in an experiment, conduct a survey, or use a computer or other types of equipment. They may have to identify which aspects of the task are most relevant. The process or procedures that they use may be observed and be an important part of the assessment. The product that is produced may take a variety of forms, such as the construction and presentation of graphs or tables, the use of photographs or drawings, or the construction of physical models. Products may

be developed over the course of several days and include opportunities for revision or modification. This freedom enables students to demonstrate their ability to select, organize, integrate, and evaluate information and ideas. The price of these gains includes the loss of efficiency, possible loss of breadth of coverage of the content domain, and greater difficulty in rating performance:

EXAMPLES

Prepare and deliver a speech to persuade people to take actions to protect the environment.

Hog is a game played with dice. The goal is to get the largest possible score. You may roll any number of dice out of a large cup. If none of the numbers is a 1, then the score for the roll is the sum of the numbers rolled. If a 1 is obtained on any of the dice, the score for the roll is zero. What number of dice do you think it best to roll? Defend your decision (Mathematical Sciences Education Board, 1993).

Write a computer program in BASIC that will sort a list of words alphabetically.

Design and carry out an investigation to estimate the acceleration, a, of a falling object such as a baseball. Describe the procedure used, present the data collected and analyzed, and state your conclusions.

Read an abridged version of the Lincoln–Douglas debates. Imagine that you were living then and heard the debates. Write a letter to a friend explaining the historical issues addressed and their importance in terms of what you know about the problems facing the nation at the time of the debates (Baker, Aschbacher, Niemi, & Sato, 1992).

Performance assessments require students to demonstrate skills by actually performing. They involve doing rather than just knowing about, and there are sometimes important differences between the two. For example, a guitar player may know which frets to press the strings against for a particular chord without being able to perform the task smoothly to produce the desired sound. Similarly, a computer programmer may know the function of various needed commands without being able to produce a correctly working program to perform a specific task, or a science student may know the parts and functions of an instrument without being able to use it properly to obtain the information needed to solve a problem. Performance assessments are needed to observe and evaluate such skills. They also communicate the message that actual performance is important.

A performance assessment task used in the 1996 NAEP Science Assessment at grade 4 is shown in Figure 11.2. As can be seen, this task requires students to do simple manipulations, to measure and record the outcomes of placing the pencil and thumbtack in the different bottles of water, to draw conclusions about the "mystery water," and to make predictions about the effects of adding salt to a solution. In this example, the manipulations, observations, and measurements are relatively simple, but these basic skills are critical in many settings and are not well assessed in a purely paper-and-pencil assessment.

The effective use of performance assessments requires careful attention to task selection and to the ways performances will be scored. Care needs to be taken in the identification of the complex skills we want to measure, in the construction of tasks that will require students to demonstrate those skills, and in the evaluation of the resulting process and/or product. Without careful attention to these aspects of the assessment, it is unlikely

FLOATING PENCIL

Using a Pencil to Test Fresh and Salt Water

You have been given a bag with some things in it that you will work with during the next 20 minutes. Take all of the things out of the bag and put them on your desk. Now look at the picture below. Do you have everything that is shown in the picture? If you are missing anything, raise your hand and you will be given the things you need.

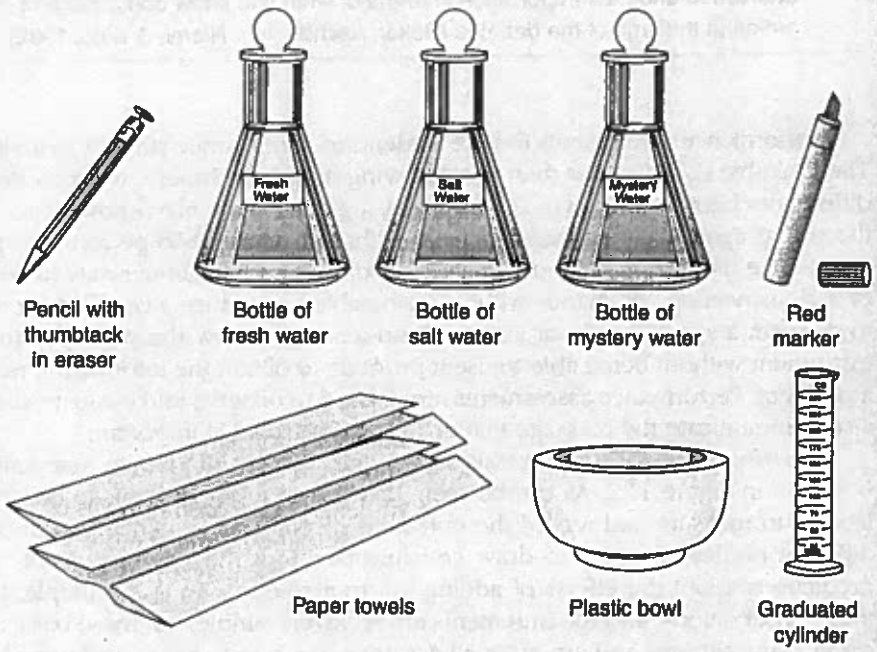


Figure 11.2
Example of hands-on science performance assessment task used at grade 4 in the 1996 National Assessment of Educational Progress

Source: NAEP 1996 Science: Report Card for the Nation and the States by O'Sullivan, C. Y., Reese, C. M., & Mazzeo, J. (1997). Washington, DC: U.S. Department of Education. Available at http://www.ed.gov/NCES/naep

that the effort will yield adequately reliable or valid measures of the complex skills that are being sought.

As the name suggests, performance assessments measure the ability of students to perform tasks that correspond to important instructional objectives. Restricted performance tasks generally focus on specific skills (e.g., reading a passage aloud). Extended performance tasks are more likely to involve problem solving and the integration of a variety of skills and understandings. A comparison of the types of complex learning outcomes measured by each of these types of performance tasks is presented in Table 11.1.

Table 11.1
Types of performance tasks

Type of Task	*Examples of Complex Learning Outcomes That Can Be Measured*
Restricted-response performance task	Ability to • read aloud • ask directions in a foreign language • construct a graph • use a scientific instrument • type a letter
Extended-response performance task	Ability to • build a model • collect, analyze, and evaluate data • organize ideas, create visuals, and make an integrated oral presentation • create a painting or perform with a musical instrument • repair an engine • write a creative short story

ADVANTAGES AND LIMITATIONS OF PERFORMANCE ASSESSMENTS

Advantages

A major advantage of performance assessments is that they can clearly communicate instructional goals that involve complex performances in natural settings in and outside of school. By using tasks that require performances that correspond as closely as is feasible to major instructional objectives, they provide instructional targets and thereby can encourage the development of complex understandings and skills. Often, performance assessment tasks are indistinguishable from good instructional activities.

A second advantage of performance assessments is that they can measure complex learning outcomes that cannot be measured by other means. As has already been stated, knowing how to do something is not the same as being able to do it, much less do it well. Thus, a paper-and-pencil test that measures what a student knows about effective public speaking, for example, does not provide a measure of the student's ability to deliver an effective speech.

A third advantage of performance assessments is that they provide a means of assessing *process* or procedure as well as the *product* that results from performing a task. For example, by observing students while they are conducting a laboratory experiment, strengths and weaknesses in the use of equipment and in technique can be assessed, as can success in completing the experiment and the strength of reasoning provided to support conclusions.

A fourth advantage of performance assessments is that they implement approaches that are suggested by modern learning theory. Rather than viewing students as recipients of discrete bits of knowledge, modern learning theory conceives of students as active participants in the construction of meaning. According to this view, new information must be actively transformed and integrated with a student's prior knowledge. High-quality performance-based assessments take student background knowledge into account and engage students in the active construction of meaning.

Limitations

The most commonly cited limitations of performance assessments parallel those cited for essay questions. Unreliability of ratings of performances across teachers or across time for the same teacher is clearly a limitation. Careful attention to the learning outcomes that the task is intended to assess and to the scoring rubrics that will be used in rating the performances is required both at the time tasks are developed and at the time performances are rated to minimize this limitation. Although the judgmental scoring of complex performances will always include some uncontrollable variations, the scoring reliability, the comparability of scores assigned to the performances of different students, and hence the fairness of the assessment can be greatly increased by clearly defining the outcomes to be measured, properly framing the tasks, and carefully defining and following rubrics for scoring performances.

Another limitation of extended performance assessments is their time-consuming nature. Because a substantial amount of time may be required to allow students to have an adequate opportunity to perform each task, relatively few extended performance assessments can be obtained within a reasonable amount of time. There is considerable evidence that performance on one task provides only a relatively weak basis for generalizing to performances on other tasks intended to assess common or related learning outcomes. Thus, solid generalization to a larger domain of outcomes requires the use of multiple tasks. Overcoming the limitation of weak generalization of performance across tasks requires the accumulation of information from performances on different tasks during the course of the year. Justification for the devotion of the required amount of instructional time to the assessments requires that the tasks provide students with good learning opportunities as well as providing assessment results.

SUGGESTIONS FOR CONSTRUCTING PERFORMANCE TASKS

The development of high-quality performance assessments that effectively measure complex learning outcomes requires attention to task development and to the ways in which performances are scored. We begin with a consideration of ways to improve the development of tasks and then suggest ways to improve scoring.

1. Focus on learning outcomes that require complex cognitive skills and student performances. It is important that tasks be interesting, but that is not sufficient. Tasks

need to be developed or selected in light of important learning outcomes. Since performance-based tasks generally require a substantial investment of student time, they should be used primarily to assess learning outcomes that are not adequately measured by less time-consuming approaches.

2. Select or develop tasks that represent both the content and the skills that are central to important learning outcomes. Current conceptions of learning stress the interdependence of content and skills. Problem solving in one subject-matter area is not the same as it is in another area. Debating a political issue in social studies is different than debating the effectiveness of a piece of literature. In each case, the content and process are interdependent. Thus, it is important to specify the range of content and resources students can use in performing a task. Past class assignments provide one natural basis for specifying content, but for many tasks it will be desirable to allow students the opportunity to do additional research to expand their knowledge base. In any event, the specification of assumed content understandings is critical to ensuring that a task functions as intended.

3. Minimize the dependence of task performance on skills that are irrelevant to the intended purpose of the assessment task. The key here is to focus on the intention of the assessment. Although both the ability to read complicated texts and the ability to communicate clearly are important learning outcomes, they are not necessarily the intent of a particular assessment. Reading ability, for example, might be irrelevant for an assessment that is intended to measure a student's ability to use mathematics to solve a practical problem (e.g., determine how much and what type of lumber to buy to build a clubhouse with specified features). However, if the task is presented in a way that requires substantial reading, then this factor may add to task difficulty for some students but not for others and thereby reduce the validity of the intended interpretation of the results. This irrelevant source of difficulty would also undermine the fairness of the assessment. On the other hand, writing skills might be an intended part of a mathematics task where a goal of the assessment was to measure a student's ability to communicate mathematical reasoning and results.

4. Provide the necessary scaffolding for students to be able to understand the task and what is expected. Challenging tasks often involve ambiguities and require students to experiment, gather information, formulate hypotheses, and evaluate their own progress in solving a problem. However, problems cannot be solved in a vacuum. Students need to have the prior knowledge and skills required to address the problem. These prerequisites can be a natural outcome of prior instruction or may be built in to the task. Preassessment activities, for example, can be used not only to introduce a task but also to ensure that students have the prior knowledge essential for the task and are familiar with the materials or equipment that they need to use. It is important to ask; What prior knowledge and skills are assumed in order to perform the task?

5. Construct task directions so that the student's task is clearly indicated. Vague directions can lead to such a diverse array of performances that it becomes impossible to rate them in a fair or reliable fashion. By design, many performance-based tasks give students a substantial degree of freedom to explore, approach problems in different ways, and come up with novel solutions. Such intended task characteristics, however, are not an excuse for vague directions. In the task shown in Figure 11.3, students need to experiment

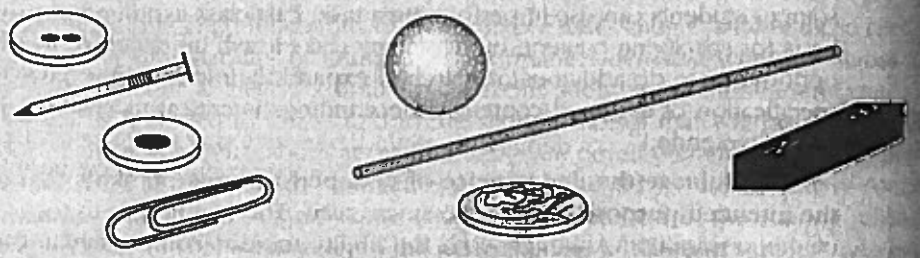
Magnet

Task Descriptor

To use a magnet to identify magnetic and nonmagnetic items and then to explain the difference between them.

Equipment/Material

A magnet and the following seven objects: plastic button, iron or steel washer, steel paper clip, iron nail, glass marble, plastic rod, and copper coin.



Student Instructions

Test the objects with the magnet and divide them into two groups. List the objects in the two groups and explain what makes the objects in the two groups different.

Scoring Scheme

Credit was given for grouping the objects correctly. Four categories of explanations were recorded: namely, that one group was made of iron or steel, that one group was attracted by the magnet, that one group was made of iron and steel and was attracted by the magnet, and any other explanation.

Figure 11.3
Example of performance assessment task in science

Source: Performance Assessment: An International Experiment by B. M. Sample, 1992, Princeton, NJ: Educational Testing Service, Report No. 22-Caep-06. Copyright 1992 by Educational Testing Service. Reprinted by permission.

and decide on the placement of objects into categories on their own. They also have to construct an explanation for the classification they provide. But the task of using the magnet to test the items, the classification of objects into two categories, and the need to explain the difference between the objects in the two categories are made explicit.

6. Clearly communicate performance expectations in terms of the scoring rubrics by which the performances will be judged. Specifying the criteria to be used in rating performance helps clarify task expectations for a student. Explaining the criteria that will be used in rating performances not only provides students with guidance on how to focus their efforts but also helps convey priorities for learning outcomes.

Listing attributes such as appropriate symbol use, accuracy of information and scale, and ease with which the map can be read makes the rating criteria explicit. It also highlights the learning outcomes that are considered important for the task in the following example:

EXAMPLE Construct a weather map. Your map will be evaluated for accuracy of information and scale, for appropriate use of symbols, and for the ease with which it can be read.

PERFORMANCE CRITERIA

Richard Stiggins (1987) has persuasively argued that the specification of performance criteria is the most important aspect of developing effective performance assessments. He suggests imagining the feedback that would be provided to a student who performed poorly before the task is administered. His rationale for focusing on the criteria to be used is straightforward: "If you do not have a clear sense of the key dimensions of sound performance—a vision of poor and outstanding performance—you can neither teach students to perform nor evaluate their performance."

The criteria to be used in judging student performance are critical for reliable, fair, and valid assessment, and the specification of the criteria should begin at the time the tasks are being selected or developed. Both the teacher and the student need to understand the criteria that will be used to judge performance. As was just noted, criteria help clarify the task expectations for students, and they communicate learning goals and standards. In addition, they guide the judgment process in ways that enhance reliability, fair treatment of each performance, and the validity of conclusions about each student's achievement.

The two main ways of guiding judgments of both the process used in performing a task and any product resulting from that performance are *scoring rubrics/rating scales* and *checklists*. We begin with rating scales and then turn to a consideration of checklists.

SCORING RUBRICS AND RATING SCALES

As was discussed in Chapter 10, a scoring rubric is a set of guidelines for the application of performance criteria to the responses and performance of students. A scoring rubric typically consists of verbal descriptions of performance or aspects of student responses that distinguish between advanced, proficient, partially proficient, and beginning levels of performance. Both analytic (Table 10.2) and holistic (Table 10.3) scoring rubrics were illustrated in Chapter 10.

The analytic scoring rubric requires the identification of different dimensions or characteristics of performance that are rated separately. For example, a mathematics task might be rated in terms of the accuracy of the calculations and the clarity of the explanation. A written report on the results of a science experiment might be rated on factual accuracy, quality of analysis, and the degree to which conclusions were justified. A literary criticism might be rated for organization, quality of ideas, clarity of expression, and mechanics.

An oral presentation might be rated both for the substantive quality of the report and for the effectiveness of the presentation.

A holistic rubric provides descriptions of different levels of overall performance. Holistic rubrics are efficient and correspond more directly to global judgments required in the assignment of grades, but they do not provide students with specific feedback about the strengths and weaknesses of their performance as is provided by analytic rubrics.

Rating scales are often limited to making quality judgments (e.g., excellent, good, fair, or poor) or scaled frequency judgments (e.g., always, frequently, sometimes, or never) for each level. As is illustrated in some of the following examples, however, the distinction between scoring rubrics and rating scales is often blurred by adding the descriptions of a rubric to the judgmental qualities of a rating scale.

As is illustrated in Figure 11.4, a scoring rubric may include a rating scale (e.g., excellent, good, and so on) but may also provide descriptions of characteristics or performance corresponding to each point on the scale. A scoring rubric makes explicit the criteria that are used to rate performance. Generic scoring rubrics are available that can be readily adapted for use in rating performance on a variety of tasks. Generic scoring rubrics such as the one shown in Figure 11.4, provide a useful starting place for many assessments. The distinctions between the levels can be made more specific by considering the specific task and likely features that would distinguish between exemplary performance

Quality of Explanation

6 = Excellent explanation (complete, clear, unambiguous)
5 = Good explanation (reasonably clear and complete)
4 = Acceptable explanation (problem completed but may contain minor flaws in explanation)
3 = Needs improvement (on the right track but may contain serious flaws; demonstrates only partial understanding)
2 = Incorrect or inadequate explanation (shows lack of understanding of problem)
1 = Incorrect without attempt at explanation

Separate Ratings of Answer and Explanation

Answer
4 = Correct
3 = Almost correct or partially correct
2 = Incorrect but reasonable attempt
1 = Incorrect with no relationship to the problem
0 = No answer

Explanation
4 = Complete, clear, logical
3 = Essentially correct but incomplete or not entirely clear
2 = Vague or unclear but with redeeming features
1 = Irrelevant, incorrect, or no explanation

Figure 11.4
Examples of generalized scoring rubrics for mathematics problems

and competent performance or between satisfactory performance with minor flaws and performance that has serious flaws. For example, lists of minor and major flaws might be constructed for a specific task. In a similar fashion, common misconceptions that are anticipated in response to a particular task might be listed.

The number of levels and the verbal descriptions used to guide the scoring may vary from situation to situation. For the hands-on science task involving the floating pencil shown in Figure 11.2, for example, the separate scoring rubrics were used for each part of the response. For the part of the task where the student was supposed to identify the mystery water and explain how they could "tell what the mystery water is," student responses were scored using a rubric with three levels:

Complete: Student stated that "the mystery water was fresh water and gave a satisfactory explanation that referred to observations made doing the hands-on task" (O'Sullivan, Reese, & Mazzeo, 1997, p. 44).
Partial: Student stated that the water was fresh but did not support the choice with direct reference to observations from the hands-on task.
Incorrect: Student gave the wrong answer or gave contradictory explanation for the choice of the correct answer of fresh water.

EXAMPLE **TASK**

First-grade children are asked to arrange four pictures of scenery in the order of the seasons by pasting them in four boxes and printing the name of each season in the box.

SCORING RUBRIC

2 points: Student arranges pictures in correct order and names of season are correctly matched to seasons.
1 point: Student begins the task but does not have a fully complete answer with four pictures in order, labeled with correct seasons.
0 points: Student does not respond appropriately.

Task and scoring guide adapted from part of a Utah State Office of Education set of assessment tasks called *Weathercaster's Helper* for first-grade students (Regional Educational Laboratories, 1998).

Scoring rubrics for hands-on tasks may include multiple dimensions, each of which focuses on a particular aspect of the process of carrying out the task. For example, in an elementary school science task used by Shavelson, Baxter, and Pine (1991) and Shavelson, Baxter, and Gao (1993), students were asked to determine which of several paper towels absorbed the most water. The scoring rubric records the method used to get the towel wet, the saturation of each towel, the procedure used to measure the amount of water absorbed, the care in measurement, and the accuracy of the result.

Rating scales provide a flexible way of converting information about one or more characteristics of a performance (e.g., overall quality, adequacy of measurement, and appropriateness of summary of results). Typically, a rating scale consists of a set of

characteristics or qualities to be judged and some type of scale for indicating the degree to which each attribute is present. The rating form itself is merely a reporting device. Its value in appraising the learning and development of students depends largely on the care with which it is prepared and the appropriateness with which it is used. As with other assessment instruments, it should be constructed in accordance with the learning outcomes to be assessed, and its use should be confined to those areas in which there is a sufficient opportunity to make the necessary observations. If these two principles are properly applied, a rating scale will serve several important assessment functions: (a) it will direct observation toward specific aspects of performance, (b) it will provide a common frame of reference for rating the performance of all students on the same set of characteristics, and (c) it will provide a convenient method for recording the observer's judgments.

Types of Rating Scales

Rating scales may take many forms, but most of them belong to one of the types described next. Each type is illustrated by using two dimensions from a scale for rating contributions to class discussion.

Numerical Rating Scale. One of the simplest types of rating scales is that in which the rater checks or circles a number to indicate the degree to which a characteristic is present. Typically, each of a series of numbers is given a verbal description that remains constant from one characteristic to another. In some cases, it is merely indicated that the larger the number is high, one is low, and the other numbers represent intermediate values.

The numerical rating scale is useful when the characteristics or qualities to be rated can be classified into a limited number of categories and there is general agreement concerning the category represented by each number. As commonly used, however, the numbers are only vaguely defined, so the interpretation and use of the scale vary:

EXAMPLE *Directions:* Indicate the degree to which this student contributes to a group problem-solving task by circling the appropriate number. The numbers represent the following values: 4—consistently appropriate and effective; 3—generally appropriate and effective; 2—needs improvement, may wander from topic; and 1—unsatisfactory (disruptive or off topic).

1. To what extent does the student participate in group discussions?

   1 2 3 4

2. To what extent are the comments related to the topic under discussion?

   1 2 3 4

Graphic Rating Scale. The distinguishing feature of the graphic rating scale is that each characteristic is followed by a horizontal line. The rating is made by placing a check on the line. A set of categories identifies specific positions along the line, but the rater is free to check between these points:

EXAMPLE *Directions:* Indicate the degree to which this student contributes to a group problem-solving task by placing an *X* anywhere along the horizontal line under each item.

1. To what extent does the student participate in group discussion?

never seldom occasionally frequently always

2. To what extent are the comments related to the topic under discussion?

never seldom occasionally frequently always

The scale shown in this example uses the same set of categories for each characteristic and is commonly referred to as a *constant-alternatives scale.* When these categories vary from one characteristic to another, the scale is called, quite logically, a *changing-alternatives scale.*

Although the line in the graphic rating scale makes it possible to rate at intermediate points, using single words to identify the categories has no great advantage over the use of numbers. There is little agreement among raters concerning the meaning of such terms as *seldom, occasionally,* and *frequently.* What is needed are descriptions of performances that indicate more specifically how students behave who possess various degrees of the characteristic being rated.

Descriptive Graphic Rating Scale. The descriptive graphic rating scale uses descriptive phrases to identify the points on a graphic scale. The descriptions are thumbnail sketches of how students behave at different steps along the scale. In some scales, only the center and end positions are defined. In others, a descriptive phrase is placed beneath each point. A space for comments is also frequently provided to enable the rater to clarify the rating:

EXAMPLE *Directions:* Make your ratings on each of the following characteristics by placing an *X* anywhere along the horizontal line under each item. In the space for comments, include anything that helps clarify your rating.

1. To what extent does the student participate in group discussions?

Never participiapates; quiet, passive

Particiapates as much as other group members

Particiapates more than any other group member

Comment:

2. To what extent are the comments related to the topic under discussion?

Comments ramble, distracts from topic

Comments usually pertinent, occasionally wanders from topic

Comments are always related to topic

Comment:

The descriptive graphic rating scale is generally the most satisfactory for school use. It explains to both the teacher and the student the types of performance that represent different degrees of progress toward desired learning outcomes. In well-written rubrics the top level of description actually is the desired learning outcome or at least communicates what good work is intended to look like. The more specific performance descriptions also contribute to greater objectivity and accuracy during the rating process. To aid scoring, numbers also may be added to each position on the scale.

Uses of Rating Scales

Rating scales can be used to assess a wide variety of learning outcomes and aspects of development. As a matter of convenience, these uses may be classified into two assessment areas: (a) process or procedure and (b) product.

Process or Procedure Assessment. In many areas, achievement is expressed specifically through the student's performance. Examples include the ability to give a speech, manipulate laboratory equipment, work effectively in a group, sing, play a musical instrument, and perform various physical feats. Such activities do not result in a product that can be assessed, and short-answer or fixed-response tests are generally inadequate. Consequently, the process or procedures used in the performance itself must be observed and judged.

Rating scales are especially useful in assessing process or procedures because they focus on the same aspects of performance in all students and have a common scale on which to record our judgments. If the rating form has been prepared in terms of specific learning outcomes, it also serves as an excellent teaching device. The dimensions and behavior descriptions used in the scale show the student the type of performance desired.

Two items from a typical rating scale for assessing a speech are presented in Figure 11.5. The first part of the form is devoted to the content of the speech and how well it is organized. The second part is concerned with aspects of delivery, such as gestures, posture, appearance, eye contact, voice, and enunciation. In developing such a scale, a teacher must, of course, include those characteristics that are most appropriate for the type of speaking ability to be assessed and for the age level of the student to be judged.

Product Assessment. When student performance results in some type of product, it is frequently more desirable to judge the product rather than the process or procedure. The ability to write a theme, for example, is best assessed by judging the quality of the theme itself. Little is to be learned by observing the student's performance. In some areas, however, such as conducting work in the laboratory, and woodworking, it might be more desirable to rate procedures during the early phase of learning and products later, after the basic skills have been mastered. In any event, product rating can provide assessment information in many areas. In addition to those already mentioned, it is useful in assessing such things as handwriting, drawings, maps, graphs, notebooks, term papers, book reports, results of laboratory experiments, and objects made in vocational courses.

A rating scale serves somewhat the same purpose in product assessment that it does in process assessment. It helps us judge the products of all students in terms of the same characteristics, and it emphasizes to the students those qualities desired in a superior product.

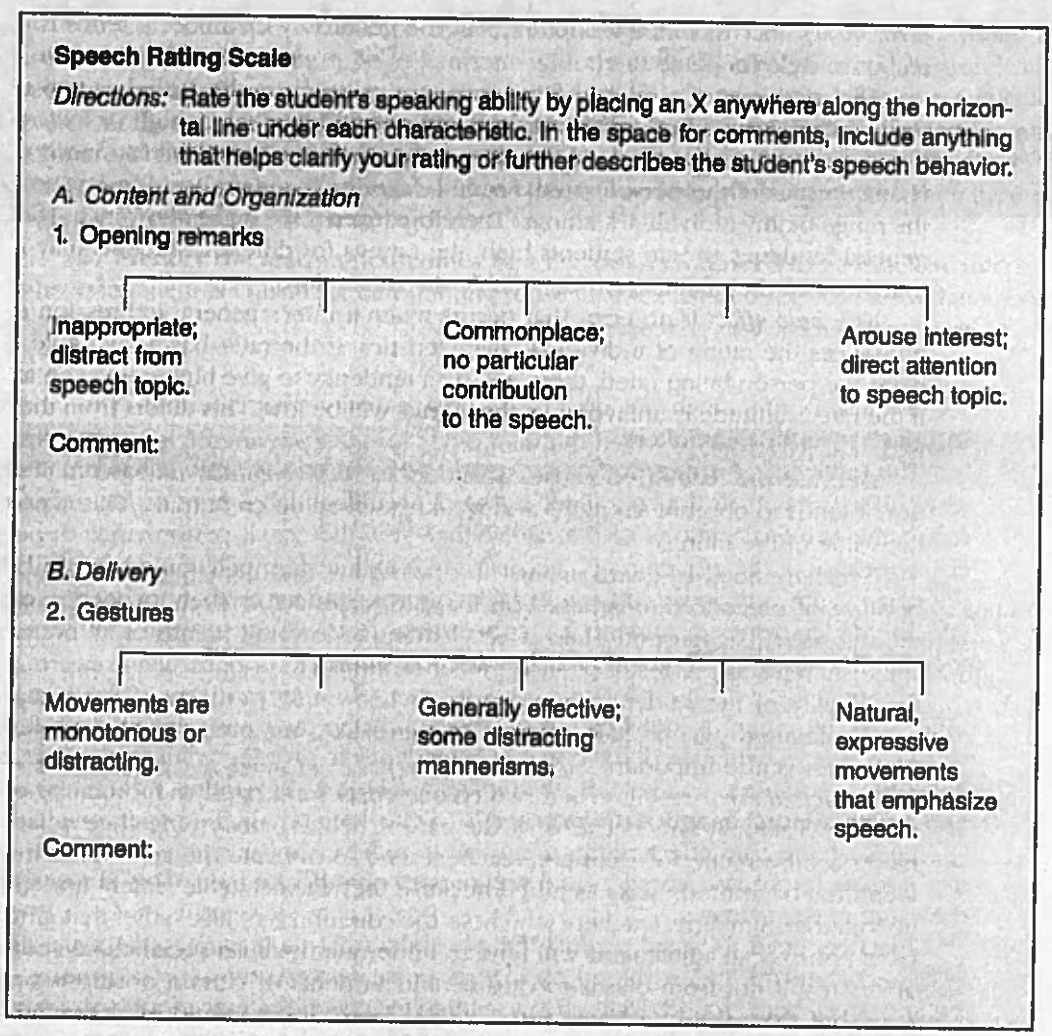


Figure 11.5
Sample items from speech rating scale

Common Errors in Rating

Certain types of errors occur so often in ratings that special efforts are needed to counteract them. These include (a) personal bias, (b) halo effect, and (c) logical errors.

Personal bias errors occur when there is a general tendency to rate all individuals at approximately the same position on the scale. Some raters tend to use the high end of the scale only, which is referred to as the *generosity error.* Occurring less frequently (but persistently for some raters) is the *severity error,* in which the lower end of the scale is favored. A third type of constant response is shown by the rater who avoids both extremes of the scale and tends to rate everyone as average. This is called the *central tendency*

error. It also occurs much less often than the generosity error, but it tends to be a fixed-response style for some raters.

The tendency of a rater to favor a certain position on the scale has two undesirable results. First, it puts in doubt a single rating of an individual. A high or low rating might reflect the personal outlook of the rater rather than the actual performance or personal characteristics of the person rated. Second, favoring a certain position on the scale limits the range of any individual's ratings. Therefore, even if we make allowances for a teacher's general tendency to rate students high, the ratings for different students may be so close together that they fail to provide reliable discriminations.

The *halo effect* is an error that occurs when a rater's general impression of a person influences the rating of individual characteristics. If the rater has a favorable attitude toward the person being rated, there will be a tendency to give high ratings on all traits, but if the rater's attitude is unfavorable, the ratings will be low. This differs from the generosity and severity errors, in which the rater tends to rate everyone high or everyone low.

Because the halo effect causes a student to receive similar ratings on all characteristics, it tends to obscure strengths and weaknesses on different traits. This obviously limits the value of the ratings.

Teachers need to guard against the possibility that their ratings might be distorted because of preconceptions based on inappropriate factors such as gender, race, ethnicity, and social background. Halo effects leading to lowered ratings of all performances of some students as the result of such preconceptions are of particular concern. Concealing the identity of the student where feasible when rating products of performance is one good safeguard against halo effects. Awareness of our own personal preferences and prejudices is also important.

A *logical error* results when two characteristics are rated as more alike or less alike than they actually are because of the rater's beliefs concerning their relationship. In rating achievement, for example, teachers tend to overrate the achievement of students identified by aptitude tests as gifted because they expect achievement and giftedness to go together. Similarly, teachers who hold the common but false belief that gifted students have poor social adjustment will tend to underrate them on social characteristics. These errors result not from biases toward certain students or certain positions on the rating scale but from the rater's assumption of a more direct relationship among traits than actually exists.

The various types of errors that appear in ratings are rather disconcerting to the classroom teacher who must depend on rating scales for assessing certain aspects of learning and development. Fortunately, however, the errors can be markedly reduced by proper design and use.

Principles of Effective Rating

The improvement of ratings requires careful attention to selection of the characteristics to be rated, design of the rating form, and conditions under which the ratings are obtained. The following principles summarize the most important considerations in these areas. Because the descriptive graphic rating scale is the most generally useful form for school purposes, the principles are directed toward the construction and use of this type of rating scale.

1. Characteristics should be educationally significant. Rating scales, like other assessment instruments, must be in harmony with the school's objectives and desired learning outcomes. Thus, when constructing or selecting a rating scale, the best guide for determining what characteristics are most significant is the list of intended learning outcomes. When these have been clearly stated in performance terms, it is often simply a matter of selecting those that can be most effectively assessed by ratings and then modifying the statements to fit the rating format (see the "Guidelines" box).

2. Identify the learning outcomes that the task is intended to assess. The intent of the assessment is critical for determining those characteristics of performance that should

GUIDELINES

Preparing Rating Scales

The same basic principle guiding the construction of test items should be followed in preparing rating scales. That is, the instrument should be designed to measure the student performance described in the instructional objectives. Let us assume, for example, that a science teacher has listed the following outcomes as evidence of skill in one phase of laboratory performance.

Demonstrates Effective Use of Laboratory Equipment

1. Selects proper equipment for a given experiment.
2. Sets up equipment quickly and correctly.
3. Manipulates equipment as needed during the experiment.
4. Measures accurately with each measuring device.
5. Follows safety rules when using equipment.
6. Cleans and returns equipment to its proper place.
7. Interprets the results of the experiment appropriately.
8. Integrates results with other knowledge in drawing conclusions.

This list of intended outcomes can then serve as the basis for preparing a rating scale to assess skill in using laboratory equipment. Each item in the list becomes an item in the rating form by simply adding some basis for recording degrees of effectiveness, as follows:

Selecting Laboratory Equipment

1	2	3	4	5
Cannot select equipment without help		Inconsistent in selecting proper equipment		Consistently selects proper equipment

The same procedure is followed when rating an educational product (e.g., theme, graph, painting, or shop and home economics projects). The characteristics of a good product are listed, and these then become the items in the rating scale. The instrument itself is simply a convenient form for recording observations and judgments concerning the extent to which students are meeting the criteria specified in the objectives.

determine the ratings. Clear identification of the learning outcomes helps establish priorities for rating, distinguish levels of performance in terms of learning outcomes, and reduce dependence on factors that are irrelevant to the intent of the assessment. When there are multiple learning outcomes associated with the task, separate ratings corresponding to each outcome may be desirable and can enhance the value of the formative feedback that is provided to students.

3. Characteristics should be directly observable. There are two considerations involved in direct observation. First, the characteristics should be limited to those that occur in school situations so the teacher has an opportunity to observe them. Second, they should be characteristics that are clearly visible to an observer. Overt behaviors, such as participation in classroom discussion, clear enunciation, and use of facts to support an argument, can be readily observed and reliably rated. However, less tangible types of behavior, such as interest in history, attitude toward literature, and amount of effort expended in library research, tend to be unreliably rated because their presence must be inferred from outward signs that are indefinite, variable, and easily faked. Whenever possible, we should confine our ratings to those characteristics that can be observed and judged directly.

4. Characteristics and points on the scale should be clearly defined. Many rating errors arise from the use of vague characterizations and inadequate identification of the scale points. The brief descriptions used with the descriptive graphic rating scale help overcome this weakness. They explain both the points on the scale and each characteristic being rated. When it is infeasible or inconvenient to use a descriptive scale, as on the back of a school report card, a separate sheet of instructions can be used to provide the desired descriptions.

5. Select the type of scoring rubric that is most appropriate for the task and the purpose of the assessment. With a holistic rubric, each performance is given a single rating or score, usually on a scale with 4 to 6 points, based on an overall judgment of the quality of the performance in comparison to the criteria specified in the scoring rubric. Holistic rubrics are efficient and translate easily into grades. As already noted, however, analytic scoring rubrics have more diagnostic value because they focus attention on those aspects of performance where improvement is needed. For analytic scores to be of diagnostic value, the characteristics or dimensions being rated must be sufficiently distinct to allow each to be reliably rated and not simply be redundant reflections of the same global impression of the performance.

6. Between three and seven rating positions should be provided. The exact number of points to be designated on a particular scale is determined largely by the judgments to be made. In areas permitting only crude judgments, fewer scale positions are needed. There is usually no advantage in going beyond the 7-point scale. Only rarely can we make finer discriminations than this, and we can provide for those few situations by allowing the rater to mark between points.

7. Rate performances of all students on one task before going on to the next one. The advantages of rating all performances on one task before starting another task parallel those described for scoring all answers to an essay question before going on to the next question. It is easier to keep the scoring criteria clearly in mind and to apply them more

uniformly when considering only a single task at a time than it is when going from task to task for each student. It also reduces the likelihood that judgments of performance on one task will be contaminated by judgments of a student's performance on a preceding task. When responses of a single student to several tasks are considered one after another, there is a strong tendency for the performance on early tasks to create an expectation for performance on later tasks. Those expectations can result in more lenient or more stringent ratings of performance than would otherwise be given.

By rating one task at a time for all students before going to the next task, it is also possible to change the order in which student performances are rated. Thus, a student is not rated first or last on all tasks or right after another student who has exceptionally good performance or exceptionally bad performance on all tasks.

8. When possible, rate performances without knowledge of the student's name. This suggestion is the same as the one given for scoring answers to essay questions. Obviously, it is not possible for all types of performance (e.g., an oral presentation), but it is good practice whenever possible. It is a practice that enhances the fairness of ratings because it reduces the chances that ratings will be influenced by a halo effect rather than only by the actual performance of a student.

9. When results from a performance assessment are likely to have long-term consequences for students, ratings from several observers should be combined. The pooled ratings of several teachers will generally yield a more reliable description of student performance than that obtained from any one teacher. In averaging ratings, the personal biases of individual raters tend to cancel out one another, but there is still a need to be alert for biases that may be shared due to similarity of background and experiences of teachers doing the rating.

CHECKLISTS

A checklist is similar in appearance and use to the rating scale. The basic difference between them is in the type of judgment needed. On a rating scale, one can indicate the degree to which a characteristic is present or the frequency with which a behavior occurs. The checklist, on the other hand, calls for a simple yes–no judgment. It is basically a method of recording whether a characteristic is present or absent or whether an action was or was not taken. Obviously, a checklist should not be used when degree or frequency of occurrence is an important aspect of the appraisal.

The checklist is especially useful at the primary level, where much of the classroom assessment depends on observation rather than testing. A simple checklist for assessing the mastery of mathematics skills at the beginning primary level is shown in Figure 11.6. If the intended learning outcomes are stated as specifically as this for each learning area, a checklist can be prepared by simply adding a place to check yes or no. As with the rating scales, the stated learning outcomes specify the performance to be assessed, and the checklist is merely a convenient means of recording judgments.

Checklists are also useful in assessing those performance skills that can be divided into a series of specific actions. An example of such a checklist for the planting a tree is

Mathematics Skills Checklist

Primary Level

Directions: Circle YES or NO to indicate whether skill has been demonstrated.

YES	NO	1. Identifies numerals 0 to 10.
YES	NO	2. Counts to 10.
YES	NO	3. Groups objects into sets of 1 to 10.
YES	NO	4. Identifies basic geometric shapes (circle, square, rectangle, triangle).
YES	NO	5. Identifies coins (penny, nickel, dime).
YES	NO	6. Compares objects and identifies bigger–smaller, longer–shorter, heavier–lighter.
YES	NO	7. States ordinals for a series of 10 objects (1st, 2nd, 3rd, etc.).
YES	NO	8. Copies numerals 1 to 10.
YES	NO	9. Tells time to the half hour.
YES	NO	10. Identifies one-half of an area.

Figure 11.6
Checklist for evaluating students' mastery of beginning skills in mathematics

shown in Figure 11.7. The performance has been subdivided into a series of observable steps, and the observer simply checks whether each step was satisfactorily completed. The checklist in Figure 11.7 includes mostly those actions that are desired in a good performance. In some cases, it may be useful to add those actions that represent common errors so that they can be checked if they occur. In Figure 11.7, for example, we might add after Item 5, "Does *not* place tree on an angle not perpendicular to the ground." If the checklist is to be used by students, the incorrect actions should, of course, be clearly identified as such.

The following steps summarize the development of a checklist for assessing a procedure consisting of a series of sequential steps:

1. Identify each of the specific actions desired in the performance.
2. Add to the list those actions that represent common errors (if they are useful in the assessment, are limited in number, and can be clearly stated).
3. Arrange the desired actions (and likely errors, if used) in the approximate order in which they are expected to occur.
4. Provide a simple procedure for checking each action as it occurs (or for numbering the actions in sequence, if appropriate).

In addition to its use in assessment of process, the checklist can also be used to assess products. For this purpose, the form usually contains a list of characteristics that the finished product should possess. In assessing the product, the teacher simply checks whether each characteristic is present or absent. Before using a checklist for product assessment, you should decide whether the quality of the product can be adequately

Directions: For each item, check when the proper behavior has been followed in planting a tree.

_____ 1. Dig a hole as deep as the rootball and twice as wide.
_____ 2. Loosen up soil around the tree hole if it is too hard.
_____ 3. Remove the container from the rootball.
_____ 4. Loosen the roots so they are not knotted up.
_____ 5. Place the tree in the hole, making sure that the tree is at the right level of depth for the ground.
_____ 6. Fill the soil in around the roots and pack the soil with your hands and feet to remove any air pockets.
_____ 7. Make a small dam around the base of the tree with the leftover soil to hold in water.
_____ 8. Soak the base of the tree with water.

Figure 11.7
Checklist for correctly planting a tree

described by merely noting the presence or absence of each characteristic. If quality is more precisely indicated by noting the degree to which each characteristic is present, a rating scale should be used instead of a checklist.

In the area of personal–social development, the checklist can be a convenient method of recording evidence of growth toward specific learning outcomes. Typically, the form lists the behaviors that have been identified as representative of the outcomes to be assessed. In the area of work habits, for example, a primary teacher might list the following behaviors (to be marked yes or no):

- Follows directions
- Seeks help when needed
- Works cooperatively with others
- Waits turn in using materials
- Shares materials with others
- Tries new activities
- Completes started tasks
- Returns equipment to proper place
- Cleans work space

Although such items can be used in checklist form if only a crude appraisal is desired, they can also be used in rating scale form by recording the frequency of occurrence (e.g., always, sometimes, never).

Although we have described the individual use of checklists, rating scales, and anecdotal records (see Chapter 13), they are often used in combination when assessing student performance (see Table 11.2).

Table 11.2
Combining techniques to assess laboratory performance in science

Types of Proficiency	*Examples of Performance to Be Assessed*	*Assessment Techniques*
Knowledge of experimental procedures	Describes relevant procedures Identifies equipment and uses Criticizes defective experiments	Paper-and-pencil testing Laboratory identification tests
Skill in designing an experiment	Plans and designs an experiment to be performed	Performance assessment with focus on product (checklist)
Skill in conducting the experiment	Selects equipment Sets up equipment Conducts experiment	Performance assessment with focus on process (rating scale)
Skill in observing and recording	Describes procedures used Reports proper measurements Organizes and records results	Performance assessment (analysis of report)
Skill in interpreting results	Identifies significant relationships Identifies weaknesses in data States valid conclusions	Performance assessment and oral questioning
Work habits	Manipulates equipment effectively Completes work promptly Cleans work space	Performance assessment with focus on process (checklist)

STUDENT PARTICIPATION IN RATING

In this chapter, we have limited our discussion to rating scales and checklists used by the teacher. We purposely omitted those checklists and rating scales used as self-report techniques by students because these will be considered in the following chapter. Before closing our discussion here, however, we should point out that most of the devices used for recording the teacher's observations also can be used by students to judge their own progress. From an instructional standpoint, it is often useful to have students rate themselves (or their products) and then compare the ratings with those of the teacher. If this comparison is made during an individual conference, the teacher can explore with each student the reasons for the ratings and discuss any marked discrepancies between the two sets.

Self-rating by a student and a follow-up conference with the teacher can have many benefits. It should help the student (a) understand better the instructional objectives, (b) recognize the progress being made toward the objectives, (c) diagnose more effectively particular strengths and weaknesses, and (d) develop increased skill in self-assessment. Of special value to the teacher is the additional insight gained.

Student participation need not be limited to the use of the assessment instruments. It is also useful to have students help develop the instruments. Through class discussion

for example, they can help identify the qualities desired in a good speech or a well-written report. A list of these suggestions can then be used as a basis for constructing a rating scale or checklist. Involving students in the development of assessment devices has special instructional values. First, it directs learning by causing the students to think more carefully about the qualities to strive for in a performance or product. Second, it has a motivating effect because students tend to put forth most effort when working toward goals they have helped define.

SUMMARY

Performance tasks provide a means of assessing a variety of student skills that cannot be measured by objective tests. To name just a few of the possibilities in addition to written responses, the performances may include oral communication; collaborating with other students; the construction of models, graphs, diagrams, or maps; or the use of tools and equipment (computers, or scientific or musical instruments). Unlike objective items, both the *process* and the *product* resulting from the performance can be assessed. Because they are time consuming both for students to do and for teachers to rate, the emphasis on performance assessment should be on measuring complex achievement that cannot be measured well by objective tests.

Restricted-response tasks are more structured and require less time to administer than extended-response tasks. These features facilitate reliability and wider coverage of a content domain. Extended-response tasks are best suited to the measurement of more complex learning outcomes, such as gathering, organizing, synthesizing, evaluating, and presenting information.

Extended performance tasks underscore the importance attached to effective performance and provide an effective means of measuring significant learning outcomes. They are the only feasible approach for measuring some important learning outcomes, they allow for the assessment of process as well as product, and their emphasis on the engagement of students in the active construction of meaning is consistent with modern learning theory. Their limitations are due mainly to the unreliability of judgmental ratings and to the time-consuming nature of the tasks and rating. Careful attention to rating criteria is critical for minimizing the unreliability due to scoring. Because of the limited generalizability of performance across tasks designed to measure the same or similar learning outcomes, it is important to base decisions on evidence accumulated from several tasks.

Rating methods are a systematic procedure for obtaining and recording the observers' judgments. Of the several types of rating scales available, the descriptive graphic scale seems to be the best for school use. In rating procedures, products, and various aspects of personal–social development, certain types of errors commonly occur. These include personal bias, halo effect, and logical errors. The control of such errors is a major consideration in constructing and using rating scales. Effective ratings result when we (a) select educationally significant characteristics, (b) identify the learning outcomes that the task is intended to assess, (c) limit ratings to directly observable behavior, (d) define clearly the characteristics and the points on the scale, (e) select the most appropriate rating procedure, (f) limit the number of points on the scale, (g) rate performances of all students on

one task before going on to the next ones, (h) rate performances without knowledge of the student's name whenever possible, and (i) combine ratings from several raters when results may have long-term consequences for students.

Checklists perform somewhat the same functions as rating scales. They are used in assessing both process and products where assessment is limited to a simple present-absent judgment.

Having students help construct and use rating devices has special values from the standpoint of learning and aids in the development of self-assessment skills.

LEARNING EXERCISES

1. In an area in which you are teaching or plan to teach, identify several learning outcomes that can be best measured with performance-based assessment tasks. For each learning outcome, construct two tasks.
2. What factors should be considered in deciding whether extended performance assessment tasks are to be included in a classroom assessment? Which of the factors are most important?
3. Describe how performance assessments might be used to facilitate learning. What types of learning are most likely to be enhanced?
4. Construct a rating scale for one of the following that would be useful for assessing the effectiveness of the performance.
   a. Giving an oral report
   b. Working in the laboratory
   c. Collaborating in group work
   d. Playing some type of game
   e. Demonstrating a skill
5. Construct a rating scale or checklist for one of the following that would be useful for assessing the product.
   a. Constructing a map, chart, or graph
   b. Writing a personal or business letter
   c. Writing a theme, poem, or short story
   d. Making a drawing or painting
   e. Making a product in industrial education
6. Prepare a checklist for assessing the ability to drive an automobile. Would a rating scale be better for this purpose? What are the relative advantages of each?
7. List some of the areas of assessment in which product scales might be used for rating.

REFERENCES

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations.* Los Angeles: University of California Center for Research on Evaluation, Standards, and Student Testing.

Mathematical Sciences Education Board. (1993). *Measuring up: Prototypes for mathematics assessment.* Washington, DC: National Academy Press. See pages 141–155 for a discussion of this game and related assessment questions.

O'Sullivan, C. Y., Reese, C. M., & Mazzeo, J. (1997). NAEP 1996 Science Report Card for the Nation and the States. Washington, DC: National Center for Education Statistics. Retrieved from http://www.ed.gov/NCES/naep

Regional Educational Laboratories. (1998). *Improving classroom assessment: A toolkit for professional developers.* Available from Regional Educational Laboratories or centrally from Northwest Regional Educational Laboratory, Portland, OR. Retrieved from http://www.nwrel.org. Includes samples of performance assessments and scoring rubrics.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4,* 347–362.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33–42. Part of an instructional series of the National Council on Measurement in Education. It presents helpful guidelines for the construction of performance assessments.

FURTHER READING

Educational Testing Service. (1993). *Performance assessment sampler: A workbook.* Princeton, NJ: Educational Testing Service. This workbook presents examples of performance assessments in various subjects, with examples of student responses and scores assigned.

Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of student achievement* (9th ed.). Upper Saddle River, NJ: Pearson. Chapter 9, "Performance Assessments," and Chapter 10, "Preparing for Performance Assessments," discuss the construction and use of various types of performance assessments.

Hart, D. (1994). *Authentic assessment: A handbook for educators.* Menlo Park, CA: Addison-Wesley. Provides a variety of examples of performance-based assessments and arguments for the importance of this approach to assessment.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development. In addition to providing examples of performance assessments and guidelines for rating, the book presents a model for linking assessment and instruction.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement,* (4th ed., pp. 387–431). Westport, CT: Praeger.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992: Mathematics Report Card for the Nation and the States* (Report No. 23-ST02). Washington, DC: U.S. Department of Education.