Chapter 8
Performance Assessment

Learning Outcomes

8.1 Identify the defining characteristics of performance tests, including effective scoring techniques.

Chief Chapter Outcome

A sufficiently broad understanding of performance assessment to distinguish between accurate and inaccurate statements regarding the nature of performance tests, the identification of tasks suitable for such tests, and the scoring of students' performances

During the early 1990s, a good many educational policymakers became enamored of ***performance assessment***, which is an approach to measuring a student's status based on the way the student completes a specified task. Theoretically, of course, when the student chooses between *true* and *false* for a binary-choice item, the student is completing a task, although an obviously modest one. But the proponents of performance assessment have measurement schemes in mind that are meaningfully different from binary-choice and multiple-choice tests. Indeed, it was dissatisfaction with traditional paper-and-pencil tests that caused many educators to travel with enthusiasm down the performance-testing trail.

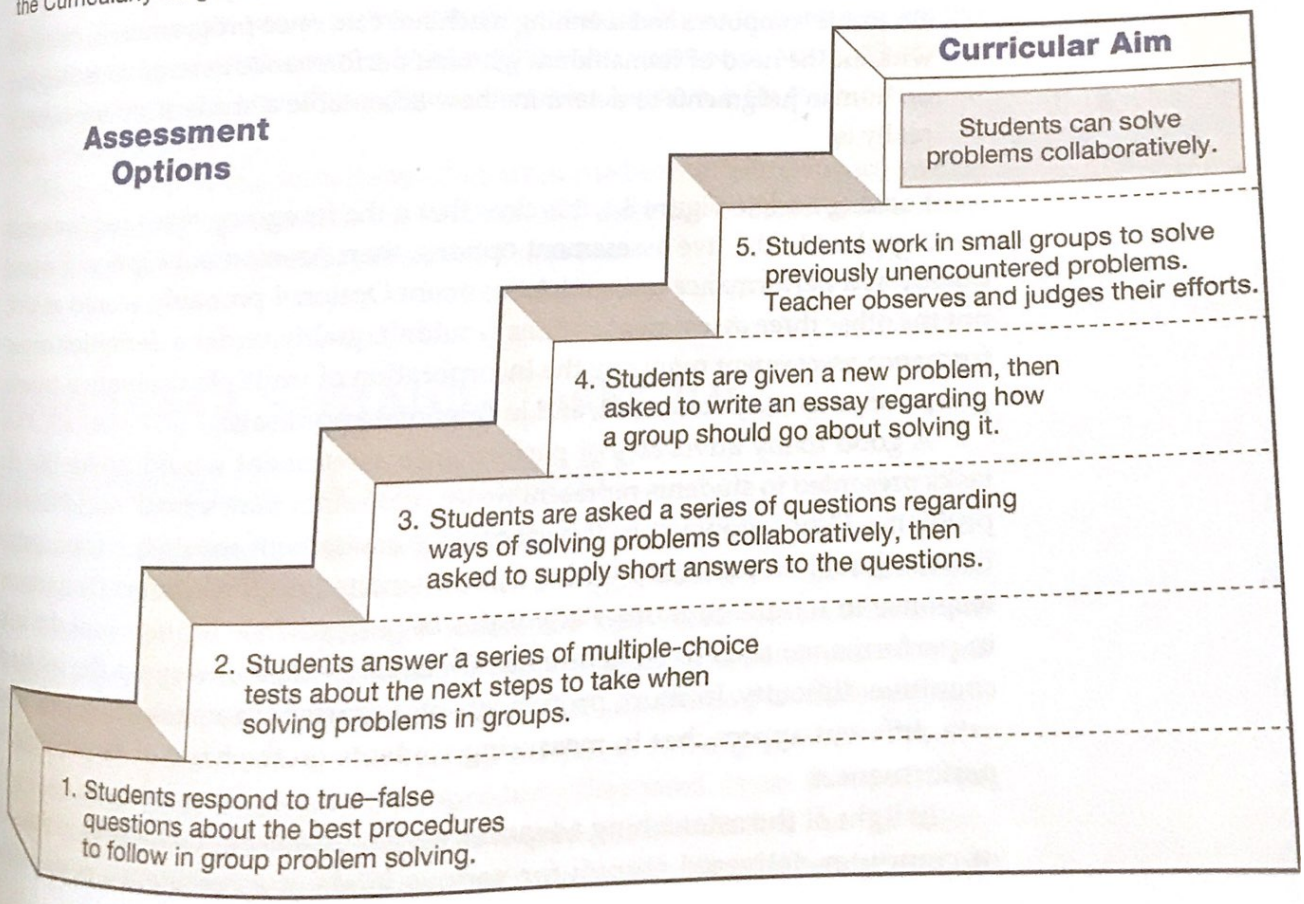
What Is a Performance Test?

Before digging into what makes performance tests tick and how you might use them in your own classroom, we'd best explore the chief attributes of such an assessment approach. Even though all educational tests require students to perform in some way, when most educators talk about performance tests, they are thinking about assessments in which the student is required to construct an original response. It may be useful for you to regard performance tests as an assessment task in which the students make products or engage in behaviors other than answering selected-response or constructed-response items. Often, an examiner (such as the teacher) *observes* the process of construction, in which case observation of the student's performance and judgment about the quality of that performance are required. Frequently, performance tests call for students to generate some sort of specified product whose quality can then be evaluated.

More than four decades ago, Fitzpatrick and Morrison (1971) observed that "there is no absolute distinction between performance tests and other classes of tests." They pointed out that the distinction between performance assessments and more conventional tests is chiefly the degree to which the examination simulates the criterion situation—that is, the extent to which the examination calls for student behaviors approximating those about which we wish to make inferences.

Suppose, for example, a teacher who had been instructing students in the process of collaborative problem solving wanted to see whether students had acquired this collaborative skill. The *inference* at issue centers on the extent to which each student has mastered the skill. The *educational decision* on the line might be whether particular students need additional instruction or, on the contrary, it's time to move on to other curricular aims. The teacher's real interest, then, is in how well students can work with other students to arrive collaboratively at solutions to problems. In Figure 8.1, you will see several assessment procedures that could be used to get a fix on a student's collaborative problem-solving skills. Note that the two selected-response assessment options (numbers 1 and 2) don't really ask students to construct anything. For the other three constructed-response assessment options (numbers 3, 4, and 5), however, there are clear differences in the degree to which

Figure 8.1 A Set of Assessment Options That Vary in the Degree to Which a Student's Task Approximates the Curricularly Targeted Behavior



the task presented to the student coincides with the class of tasks called for by the teacher's curricular aim. Assessment Option 5, for example, is obviously the closest match to the behavior called for in the curricular aim. Yet Assessment Option 4 is surely more of a "performance test" than Assessment Option 1.

It should be apparent to you, then, that different educators will be using the phrase *performance assessment* to refer to very different kinds of assessment approaches. Many teachers, for example, are willing to consider short-answer and essay tests as a form of performance assessment. In other words, those teachers essentially equate performance assessment with any form of constructed-response assessment. Other teachers establish more stringent requirements for a measurement procedure to be accurately described as a performance assessment. For example, some performance-assessment proponents contend that genuine performance assessments must exhibit at least three features:

- *Multiple evaluative criteria.* The student's performance must be judged using more than one ***evaluative criterion***. To illustrate, a student's ability to speak Spanish might be appraised on the basis of the student's accent, syntax, and vocabulary.
- *Prespecified quality standards.* Each of the evaluative criteria on which a student's performance is to be judged is clearly explicated in advance of judging the quality of the student's performance.
- *Judgmental appraisal.* Unlike the scoring of selected-response tests in which electronic computers and scanning machines can, once programmed, carry on without the need of humankind, genuine performance assessments depend on human judgments to determine how acceptable a student's performance really is.

Looking back to Figure 8.1, it is clear that if the foregoing three requirements were applied to the five assessment options, then Assessment Option 5 would qualify as a performance test, and Assessment Option 4 probably would as well, but the other three assessment options wouldn't qualify under a definition of performance assessment requiring the incorporation of multiple evaluative criteria, prespecified quality standards, and judgmental appraisals.

A good many advocates of performance assessment would prefer that the tasks presented to students represent real-world rather than school-world kinds of problems. Other proponents of performance assessment would be elated simply if more school-world measurement was constructed response rather than selected response in nature. Still other advocates of performance testing want the tasks in performance tests to be genuinely *demanding*—that is, way up the ladder of cognitive difficulty. In short, proponents of performance assessment often advocate different approaches to measuring students on the basis of those students' performances.

In light of the astonishing advances we now see every few weeks in the sorts of computer-delivered stimuli for various kinds of assessment—performance

tests surely included—the potential nature of performance-test tasks seems practically unlimited. For example, the possibility of digitally simulating a variety of authentic performance-test tasks provides developers of performance tests with an ever-increasing range of powerful performance assessments placing students in a test-generated "virtual world."

You'll sometimes encounter educators who use other phrases to describe performance assessment. For example, they may use the label ***authentic assessment*** (because the assessment tasks more closely coincide with real-life, nonschool tasks) or the label *alternative assessment* (because such assessments constitute an alternative to traditional, paper-and-pencil tests). In the next chapter, we'll be considering *portfolio assessment*, which is a particular type of performance assessment and should not be considered a synonymous descriptor for the performance-assessment approach to educational measurement.

To splash a bit or reality juice on this chapter, it may be helpful for you to recognize a real-world fact about educational performance assessment. Here it goes: Although most educators regard performance testing as an effective way to measure students' mastery of important skills instead of using many traditional testing tactics, in recent years the financial demands of such testing have rendered them nonexistent or, at best, tokenistic, in many settings. Yes, as the chapter probes the innards of this sort of educational testing, you will discover that it embodies some serious advantages, particularly its contributions to instruction. Yet, you will also see that a full-fledged reliance on performance testing for many students, such as we see in states' annual accountability tests, often renders the widespread use of performance tests prohibitively expensive. More affordable, of course, is teachers' use of this potent assessment approach for their own classroom assessments.

We now turn to the twin issues that are at the heart of performance assessments: *selecting appropriate tasks* for students and, once the students have tackled those tasks, *judging the adequacy of students' responses*.

Identifying Suitable Tasks for Performance Assessment

Performance assessment typically requires students to respond to a small number of more significant tasks, rather than to a large number of less significant tasks. Thus, rather than answering 50 multiple-choice items on a conventional chemistry examination, students who are being assessed via performance tasks may find themselves asked to perform an actual experiment in their chemistry class and then prepare a written interpretation of the experiment's results, along with an analytic critique of the procedures they used. From the chemistry teacher's perspective, instead of seeing how students respond to the 50 "mini-tasks" represented in the multiple-choice test, an estimate of each student's status

must be derived from the student's response to a single, complex task. Given the significance of each task used in a performance-testing approach to classroom assessment, it is apparent that great care must be taken in the selection of performance-assessment tasks. Generally speaking, classroom teachers will either have to (1) generate their own performance-test tasks or (2) select performance-test tasks from the increasing number of tasks currently available from educators elsewhere.

Inferences and Tasks

Consistent with the frequently asserted message of this text about classroom assessment, the chief determinants of how you assess your students are (1) the inference—that is, the interpretation—you want to make about those students

Decision Time

Grow, Plants, Grow!

Francine Floden is a third-year biology teacher in Kennedy High School. Because she has been convinced by several of her colleagues that traditional paper-and-pencil examinations fail to capture the richness of the scientific experience, Francine has decided to base most of her students' grades on a semester-long performance test. As Francine contemplates her new assessment plan, she decides that 90 percent of the students' grades will stem from the quality of their responses to the performance test's task; 10 percent of the grades will be linked to classroom participation and to a few short true–false quizzes administered throughout the semester.

The task embodied in Francine's performance test requires each student to design and conduct a 2-month experiment to study the growth of three identical plants under different conditions, and then prepare a formal scientific report describing the experiment. Although most of Francine's students carry out their experiments at home, several students use the shelves at the rear of the classroom for their experimental plants. A number of students vary the amount of light or the kind of light received by the different plants, but most students modify the nutrients given to their plants. After a few weeks of the 2-month experimental period, all of Francine's students seem to be satisfactorily under way with their experiments.

Several of the more experienced teachers in the school, however, have expressed their reservations to Francine about what they regard as "overbooking on a single assessment experience." The teachers suggested to Francine that she will be unable to draw defensible inferences about her students' true mastery of biological skills and knowledge on the basis of a single performance test. They urged her to reduce dramatically the grading weight for the performance test so that, instead, additional grade-contributing exams can also be given to the students.

Other colleagues, however, believe Francine's performance-test approach is precisely what is needed in courses such as biology. They recommended that she "stay the course" and alter "not one whit" of her new assessment strategy. (Francine was obliged to look up the technical meaning of "whit.")

If you were Francine, what would your decision be?

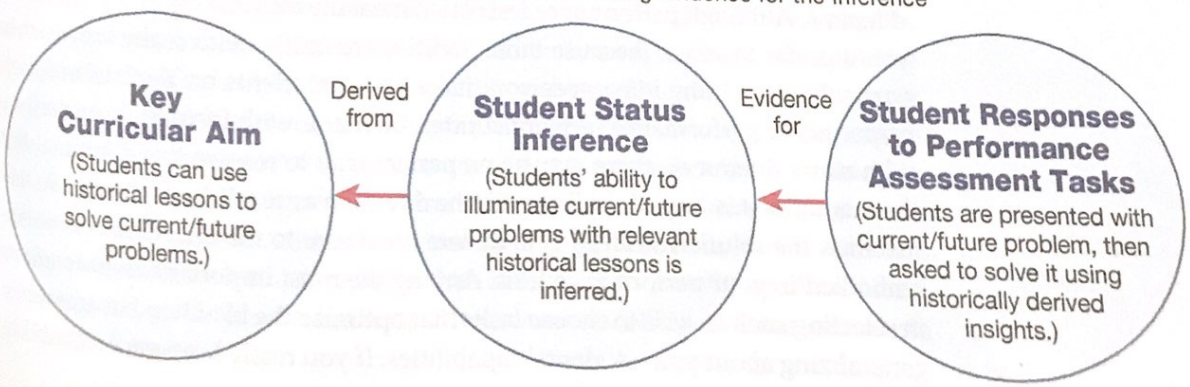
and (2) the decision that will be based on the inference. For example, suppose you're a history teacher and you've spent a summer at a lakeside cabin meditating about curricular matters (which, in one lazy setting or another, is the public's perception of how most teachers spend their summer vacations). After three months of heavy curricular thought, you have concluded that what you really want to teach your students is to apply historical lessons of the past to the solution of current and future problems that, at least to some extent, parallel the problems of the past. You have decided to abandon your week-long, 1500-item true–false final examination, which your stronger students refer to as a "measurement marathon" and your weaker students refer to by using a rich, if earthy, vocabulary. Instead of using true–false items, you are now committed to a performance-assessment strategy, and you wish to select tasks for your performance tests. You want the performance test to help you infer how well your students can draw on the lessons of the past to illuminate their approach to current and/or future problems.

In Figure 8.2, you will see a graphical depiction of the relationships among (1) a teacher's key curricular aim, (2) the inference that the teacher wishes to draw about each student, and (3) the tasks for a performance test intended to secure data to support the inference that the teacher wants to make. As you will note, the teacher's curricular aim provides the source for the inference. The assessment tasks yield the evidence needed for the teacher to arrive at defensible inferences regarding the extent to which students can solve current or future problems using historical lessons. To the degree that students have mastered the curricular aim, the teachers will reach a decision about how much more instruction, if any, is needed.

The Generalizability Dilemma

One of the most serious difficulties with performance assessment is that, because students respond to fewer tasks than would be the case with conventional paper-and-pencil testing, it is often more difficult to generalize accurately about what

Figure 8.2 Relationships among a Teacher's Key Curricular Aim, the Assessment-Based Inference Derived from the Aim, and the Performance-Assessment Tasks Providing Evidence for the Inference



skills and knowledge are possessed by the student. To illustrate, let's say you're trying to get a fix on your students' ability to multiply pairs of double-digit numbers. If, because of your instructional priorities, you can devote only a half-hour to assessment purposes, you could require the students to respond to 20 such multiplication problems in the 30 minutes available. (That's probably more problems than you'd need, but this example attempts to draw a vivid contrast for you.) From a student's responses to 20 multiplication problems, you can get a pretty fair idea about what kind of double-digit multiplier the student is. As a consequence of the student's performance on a *reasonable sample* of items representing the curricular aim, you can reasonably conclude that "Javier really knows how to multiply those sorts of problems," or "Fred really couldn't multiply double-digit multiplication problems if his life depended on it." It is because you have adequately sampled the kind of student performance (about which you wish to make an inference) that you can confidently make inferences about your students' abilities to solve similar sorts of multiplication problems.

With only a 30-minute assessment period available, however, if you moved to a more elaborate kind of performance test, you might only be able to have students respond to one big-bopper item. For example, if you presented a multiplication-focused mathematics problem involving the use of manipulatives, and wanted your students to derive an original solution and then describe it in writing, you'd be lucky if your students could finish the task in half an hour. Based on this single task, how confident would you be in making inferences about your students' abilities to perform comparable multiplication tasks?

And this, as you now see, is the rub with performance testing. Because students respond to fewer tasks, the teacher is put in a trickier spot when it comes to deriving accurate interpretations about students' abilities. If you use only one performance test, and a student does well on the test, does this mean the student *really* possesses the category of skills the test was designed to measure, or did the student just get lucky? On the other hand, if a student messes up on a single-performance test, does this signify that the student *really* doesn't possess the assessed skill, or was there a feature in this particular performance test that misled the student who, given other tasks, might have performed marvelously?

As a classroom teacher, you're faced with two horns of a classic measurement dilemma. Although performance tests often measure the kinds of student abilities you'd prefer to assess (because those abilities are in line with really worthwhile curricular aims), the inferences you make about students on the basis of their responses to performance tests must often be made with increased caution. As with many dilemmas, there may be no perfect way to resolve this dilemma. But there is, at least, a way of dealing with the dilemma as sensibly as you can. In this instance, the solution strategy is to devote great care to the selection of the tasks embodied in your performance tests. Among the most important considerations in selecting such tasks is to choose tasks that optimize the likelihood of accurately generalizing about your students' capabilities. If you really keep generalizability

at the forefront of your thoughts when you select or construct performance-test tasks, you'll be able to make the strongest possible performance-based inferences about your students' capabilities.

Factors to Consider When Evaluating Performance-Test Tasks

We've now looked at what many measurement specialists regard as the most important factor you can consider when judging potential tasks for performance assessments—*generalizability*. Let's look at a set of seven such factors you might wish to consider, whether you select a performance-test task from existing tasks or create your own performance-test tasks anew.

Evaluative Criteria for Performance-Test Tasks

- *Generalizability.* Is there a high likelihood that the students' performance on the task will generalize to comparable tasks?
- *Authenticity.* Is the task similar to what students might encounter in the real world, as opposed to encountering it only in school?
- *Multiple foci.* Does the task measure multiple instructional outcomes instead of only one?
- *Teachability.* Is the task one that students can become more proficient in as a consequence of a teacher's instructional efforts?
- *Fairness.* Is the task fair to all students—that is, does the task avoid bias based on such personal characteristics as students' gender, ethnicity, or socio-economic status?
- *Feasibility.* Is the task realistically implementable in relation to its cost, space, time, and equipment requirements?
- *Scorability.* Is the task likely to elicit student responses that can be reliably and accurately evaluated?

Whether you're developing your own tasks for performance tests or selecting such tasks from an existing collection, you may wish to apply some but not all the factors listed here. Some teachers try to apply all seven factors, although they occasionally dump the *authenticity* criterion or the *multiple foci* criterion. In some instances, for example, school tasks rather than real-world tasks might be suitable for the kinds of inferences a teacher wishes to reach, so the authenticity criterion may not be relevant. And even though it is economically advantageous to measure more than one outcome at one time, particularly considering the time and effort that goes into almost any performance test, there may be cases in which a single educational outcome is so important that it warrants a solo performance test. More often than not, though, a really good task for a performance test will satisfy most, if not all seven, of the evaluative criteria presented here.