CHAPTER 4

DESIGNING QUALITY ASSESSMENTS

CHAPTER FOCUS

This chapter answers the following guiding questions:

As a classroom teacher, how should I assess the achievement of my students?
What methods should I use, and how do I use them effectively?

From your study of this chapter, you will understand the following:

1. Quality assessments display four key design features: (1) the assessment method selected is capable of reflecting clearly defined targets; (2) they are built of quality ingredients; (3) they sample achievement sufficiently to provide valid conclusions about student mastery; and (4) they are constructed and used in ways that minimize bias and distortion in results.

2. We have four categories of assessment methods from which to choose for any particular classroom assessment situation: selected response, essay, performance assessment, and personal communication.

3. The method of choice in any particular classroom assessment context is a function of the information needs of the user and the achievement target in question.

The Assessment Options

Having introduced the idea that assessments must be valid for particular purposes (users and uses) and must be reflections of particular achievement targets, we now turn to our next validity issue: selecting proper assessment methods for a particular purpose and target. In this chapter, we explain how the various methods that we have at our disposal fit into the big picture of assessment quality (Figure 4.1).
We’re going to study four basic assessment methods, all of which are familiar to you: selected response assessments, essay assessments, performance assessments, and assessments that rely on direct personal interaction or communication with students. Each provides its own special form of evidence of student proficiency. I introduce them in this chapter, and then in Part II devote an entire chapter to each method.

Our goal as teachers and assessors is to gather valid and reliable evidence of student mastery of our valued learning targets, and to use that evidence to maximize student achievement. Quality classroom assessments display four significant design features: (1) they rely on assessment methods capable of reflecting the targets in question; (2) they are built of quality ingredients (test items, scoring schemes, etc.); (3) they sample achievement with enough tasks to lead to a confident conclusion about student mastery; and (4) they are constructed and used in ways that minimize distortion in results due to bias. We address the first feature in depth in this chapter, and introduce the other three. In Part II, we explore each of these remaining three features as they become relevant to each assessment method.
In this chapter, we will analyze how each of these four methods aligns with the four kinds of achievement targets plus dispositions discussed in Chapter 3. In essence, we will see which methods make sense with each target (can yield valid results) and which do not (will yield invalid results). We will do this by filling in the cells of Figure 4.2 with commentary on viable matches.

The artistry of classroom assessment emerges when teachers orchestrate a careful alignment among user information needs, achievement targets, and assessment methods. For example, an assessment of instrumental music proficiency is likely to look very different from an assessment of understanding of science knowledge. The former relies on the assessor to listen to and subjectively judge proficiency. The latter can be accomplished with a set of exercises scored correct or incorrect, yielding a score reflecting proficiency. Different targets, different assessment methods.

It should be self-evident that in the primary grades, before students have become confident readers and writers, selected response or essay methods cannot be used, as they require competence in those modes of communication. On the other hand, in middle school science, for example, they make perfect sense. Different settings, different methods.

Note also that an assessment of instrumental music proficiency for the purpose of planning a student’s next lesson demands a different kind of assessment from one designed to determine who receives a scholarship to the conservatory. The former requires a narrowly focused, brief assessment; the latter a much larger, more diverse sampling of proficiency. As purpose varies, so does the definition of sound assessment procedure.

<table>
<thead>
<tr>
<th></th>
<th>Selected Response</th>
<th>Essay</th>
<th>Performance Assessment</th>
<th>Personal Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reasoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Products</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dispositions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.2
A plan for matching assessment methods with achievement targets
In this chapter, we’re going to study how to pick a method that truly reflects the target you wish to assess. We also will continue to fill in more details about the idea of student involvement with each assessment method.

**Selected Response Assessment**

This category includes all of the objectively scored paper and pencil test formats. Respondents are asked a series of questions, each of which is accompanied by a range of alternative responses. Their task is to select either the correct or the best answer from among the options. The index of achievement in this instance is the number or proportion of questions answered correctly. Format options within this category include the following:

- multiple-choice items
- true/false items
- matching exercises
- short answer fill-in items

I realize that fill-in-the-blank items do require a response originating from within a respondent’s mind, but I include it in this category because it calls for a very brief answer that typically is counted right or wrong.

**Essay Assessment**

In this case, respondents are provided with exercises that call for them to prepare original written answers. Respondents might answer questions about content knowledge or provide an explanation of the solution to a complex problem. They might be asked to compare historical events, interpret scientific information, or solve open-ended math problems, where they must show and explain all their work. The examiner reads this original written response and evaluates it by applying specified scoring criteria.

Evidence of achievement is seen in the conceptual substance of the response (i.e., ideas expressed and the manner in which they are tied together). Once again, as with selected response, the student’s score is determined by the number of points attained out of a total number of points possible.

**Performance Assessment**

In this case, respondents actually carry out a specified activity under the watchful eye of an evaluator, who observes their performance or its results and judges the level of achievement demonstrated. Performance assessments can be based either on observations of respondents as they are demonstrating skills, or on the products created as a result of performing. In this sense, as with essay assessments, performance assessments consist of two parts: a performance task or assignment and a set of scoring guides.
Respondents may evidence achievement by carrying out a proper sequence of activities or by doing something in the appropriate manner. Examples include musical performance, reading aloud, communicating conversationally in a second language, or carrying out some motor activity, as in the case of physical education or dance. In this case, it is the doing that counts. The index of achievement typically is a performance rating or profile of ratings reflecting levels of quality in the performance.

Alternatively, respondents may demonstrate proficiency by creating complex achievement-related products intended to meet certain standards of quality. The product resulting from performance must exist as an entity separate from the performer, as in the case of term papers, science fair exhibits, or art and craft creations. The assessor examines the tangible product to see if those attributes of quality are indeed present. In this instance, it is not so much the process of creating that counts (although that may be evaluated, too) but rather the characteristics of the creation itself. Again in this case, the index of achievement is the rating(s) of product quality.

Personal Communication as Assessment

One of the most common ways teachers gather information about day-to-day student achievement in the classroom is to talk with them! We typically don’t think of this as assessment in the same sense as a multiple-choice test or a performance assessment. But on reflection, it will become clear to you that certain forms of personal communication definitely do provide evidence of the level of student achievement.

These forms of personal communication include questions posed and answered during instruction, interviews, conferences, conversations, listening during class discussions, and oral examinations. The examiner listens to responses and either (1) judges them right or wrong if correctness is the criterion, or (2) makes subjective judgments according to some continuum of quality. Personal communication is a very flexible means of assessment that we can bring into play literally at a moment’s notice. While it certainly is not as efficient as some other options when we must assess many students, it can probe achievement far more deeply than can the other alternatives.

Time for Reflection

Think about the assessments you have experienced in the classroom, either as a student or as a teacher, and identify an example from your personal experience of each of the four assessment methods described.

Keep the Options in Balance

As you assess in your classroom, strive to maintain a balanced perspective regarding the viability of these assessment options. For decades, one method dominated
in the United States, the objectively scored multiple-choice (selected response) test. As a result, some very important achievement targets that cannot be trans-
lated into this method simply were not assessed. This was especially true in the
context of standardized testing, where selected response methods have been used
extensively.

As we moved through the 1980s, however, we began to experience a shift in
our values regarding the dominant method. As we began to embrace a more com-
plex array of valued achievement expectations, including direct assessment of writ-
ing proficiency and some relatively more complex patterns of reasoning, we began
to make more extensive use of such alternatives as extended written response and
performance assessment.

As a result of this swing in values, however, if we aren’t careful we may go
out of balance in the other direction. By this, I mean we risk trading our prior
obsession with multiple-choice tests for a new obsession with performance
assessments. Neither method is inherently superior to the other. We can prevent
this problem only by knowing what method to use and when and how to use
each of them well.

Design Feature 1: Matching Methods with Targets

Note that three of the four assessment methods described call for students to
develop complex original responses. They require extended written responses,
demonstrate complex performance skills, create multidimensional products,
or participate in one-on-one communication, all of which take more time to
administer and certainly more time to score than, say, true/false test items.
Thus, if the amount of assessment time is held constant, selected response
assessments can provide a much larger sample of performance per unit of
assessment time.

Given this fact, you might ask, why not just use the most efficient option all the
time, selected response? The reason is that selected response assessment formats
cannot validly depict all of the kinds of achievement we expect of our students.
Different kinds of assessment methods align well with different kinds of achieve-
ment targets. We explore these relationships next.

Your objective, given a choice of methods, is to identify the most efficient (most
valid evidence per unit of time) for your specific context. As it turns out, the recipes
for creating these blends are not complicated.

As you saw in Figure 4.2, we visualize this blending by crossing the five kinds
of outcomes with the four methods to create a table depicting the various matches
of targets to methods. We may then explore the nature and practicality of the
match within each cell of this table. The result, though not a simple picture, is
both understandable and practical. Table 4.1 presents brief descriptions of the
various matches.
Important Things to Remember

As you read Table 4.1, please keep the following key points in mind:

Know Your Targets
Remember that assessments provide us with external indicators of the learner’s internal mental states (achievements). These indicators take the form of visible manifestations that we can see and count or evaluate, such as correct or incorrect responses to test items or ratings of a performance skill. In other words, because we can’t just lift the tops off students’ heads and look inside to see if math problem-solving proficiency is in there, we administer an assessment in the form of several math problems to gather evidence of proficiency from which we infer mastery of the desired achievement. Your job in selecting a method for any particular form of achievement is to choose the method that permits you to draw the most valid (defensible) inference about student learning. For this reason, I reiterate yet again, you must be a master of the targets your students are expected to hit if you are to select, develop, and use sound assessments of those targets.

Remember, We Sample Achievement to Generalize About Student Learning
Any assessment will represent a sample of all the exercises we could have posed if the assessment were infinitely long. A sound assessment relies on a sample that is systematically representative of the possibilities. We use performance on the sample to infer how much of the target students have mastered.

Our goal in assessment design is to use the most powerful assessment option we can. Power derives from the accuracy and efficiency with which a method can represent our valued standard. We always want the highest-resolution picture of that valued target we can get using the smallest possible sample of student performance; maximum information for minimum cost.

As you read Table 4.1, I hope you can see more clearly why it is crucial to understand the achievement target in order to select a proper assessment method. This cannot be overstated: Different targets require different methods.

Remember Too, This Is Just Your First Pass
This plan for aligning achievement targets and assessment methods is going to sound somewhat complex to you the first time through it. Just remember that this is only our first time! The remainder of the book is about how to make these matches work to your benefit and to the benefit of your students. All I hope for here is that the alignments described make sense to you. More practice in later chapters will help you become comfortable with these procedures.

Time for Reflection

Please take a few minutes now to study Table 4.1 and then read on. Do not read on without reading through the table.
### Table 4.1
Links between achievement targets and assessment methods

<table>
<thead>
<tr>
<th>Target to Be Assessed</th>
<th>Assessment Method</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Knowledge &amp; Understanding</strong></td>
<td>Selected Response</td>
<td>Can sample mastery of elements of knowledge</td>
</tr>
<tr>
<td></td>
<td>Essay</td>
<td>Can tap understanding of relationships among elements of knowledge</td>
</tr>
<tr>
<td></td>
<td>Performance Assessment</td>
<td>Not a good choice for this target—three other options preferred</td>
</tr>
<tr>
<td></td>
<td>Personal Communication</td>
<td>Can ask questions, evaluate answers and infer mastery, but is time-consuming option</td>
</tr>
<tr>
<td><strong>Reasoning Proficiency</strong></td>
<td>Selected Response</td>
<td>Can assess application of some patterns of reasoning</td>
</tr>
<tr>
<td></td>
<td>Essay</td>
<td>Can provide a window into reasoning proficiency</td>
</tr>
<tr>
<td></td>
<td>Performance Assessment</td>
<td>Can watch students solve some problems or examine some products and infer about reasoning proficiency</td>
</tr>
<tr>
<td></td>
<td>Personal Communication</td>
<td>Can ask student to “think aloud” or can ask followup questions to probe reasoning</td>
</tr>
<tr>
<td><strong>Performance Skills</strong></td>
<td>Selected Response</td>
<td>Can assess mastery of understandings needed for skillful performance, but cannot rely on these to tap the skill itself</td>
</tr>
<tr>
<td></td>
<td>Essay</td>
<td>Can observe and evaluate skills as they are being performed</td>
</tr>
<tr>
<td><strong>Ability to Create Products</strong></td>
<td>Selected Response</td>
<td>Can only assess mastery of the understandings needed to create quality products</td>
</tr>
<tr>
<td></td>
<td>Essay</td>
<td>Can assess mastery of knowledge needed for product development; brief essays can provide evidence of writing proficiency</td>
</tr>
<tr>
<td></td>
<td>Performance Assessment</td>
<td>Can assess (1) proficiency in carrying out steps in product development, and (2) attributes of the product itself</td>
</tr>
<tr>
<td></td>
<td>Personal Communication</td>
<td>Can probe procedural knowledge and knowledge of attributes of quality products, but not product quality</td>
</tr>
<tr>
<td><strong>Dispositions</strong></td>
<td>Selected Response</td>
<td>Questionnaire items can tap student feelings</td>
</tr>
<tr>
<td></td>
<td>Essay</td>
<td>Open-ended questionnaire items can probe dispositions</td>
</tr>
<tr>
<td></td>
<td>Performance Assessment</td>
<td>Can infer dispositions from behavior and products</td>
</tr>
<tr>
<td></td>
<td>Personal Communication</td>
<td>Can talk with students about their feelings</td>
</tr>
</tbody>
</table>
Assessing Knowledge and Understanding

Here is how our four assessment methods align with knowledge and understanding targets.

**Selected Response**

We all know that we can use selected response, objective paper and pencil tests to measure student mastery of facts, concepts, and even generalizations. Typically, these assessments tend to rely on independent items to test mastery of disconnected elements of knowledge, such as knowledge of United States history, spelling, vocabulary, earth science, and the like.

These tests are efficient in that we can administer large numbers of multiple-choice or true/false test items per unit of testing time. Thus, they permit us to sample widely and draw relatively confident generalizations from the content sampled. For this reason, when the target is knowledge mastery, selected response formats fit nicely into the resource realities of most classrooms.

But remember, even when the method aligns well with the target, things can go wrong. For instance, what if respondents confronted with a multiple-choice test can’t read? A nonreader or a student who is still learning English might actually know the material but score low because of poor reading proficiency. If we conclude that their low score means a lack of knowledge, we would be wrong. We’ll explore these and other potential sources of mismeasurement in later chapters.

**Essay**

When the domain of knowledge is defined not as elements in isolation but rather as important relationships among elements, larger concepts, and important generalizations—in other words, where the knowledge to be mastered is organized in complex ways—we can test student mastery by having them portray their knowledge using an extended written essay format. Examples of larger information chunks we might ask students to know are the causes of westward migration in U.S. history or differences among igneous, metamorphic, and sedimentary rocks.

In this case, we sample with fewer exercises, because each exercise requires longer response times, and each provides us with relatively more information than any single selected response item would.

Further, essay assessments present us with a more complex scoring challenge, and not just in terms of the time it takes. Because we must subjectively judge response quality, not just count it right or wrong, bias can creep in if we are not cautious. This requires developing and using carefully developed scoring guidelines. And remember, in this case, students also must bring writing proficiency into the assessment context. We must remain aware of the danger that students might know and understand the material but be unable to communicate it in this manner.

**Performance Assessment**

When it comes to the use of performance assessment (observation and professional judgment) to detect mastery of content knowledge, things quickly become
This match is not always a strong one. To see why, consider a brief example.

Let’s say we ask a student to complete a rather complex repair of a piece of technical equipment to determine if she understands the equipment. If the student succeeds, the equipment will work properly. So this is an instance of product-based performance assessment. We will evaluate the student’s success based on whether the piece of equipment works when she is done. If the student successfully completes the repair and the piece works properly, then obviously, she possesses the prerequisite knowledge of equipment assembly and operations needed to both identify and solve the problem. In this case, the match between performance assessment and assessment of mastery of knowledge is a strong one.

However, to understand the potential problem with this match, consider the instance in which the student fails to produce functioning equipment. Is her failure due to lack of knowledge? Or does she possess the required knowledge but cannot use it properly to identify the problem (a flaw in reasoning)? Or does the student possess the knowledge and reason productively, but fail because of inept use of repair tools (a performance skill problem)? At the time the student fails to perform successfully, we just don’t know.

In fact, we cannot know the real reason for failure unless and until we follow up the performance assessment by asking some questions to find out the cause; in short, unless we turn to one of the other assessment methods to gather more evidence. But if our initial goal simply is to determine if she has mastery of prerequisite content knowledge, why go through all this hassle? Why not just ask—that is, turn to one of the other three options from the outset?

Also understand that the purpose of the assessment represents an important consideration here. If my reason for assessing is to certify repair technicians, I don’t care why the student fails. But if I am a teacher whose job is to help students learn to perform, unless I know why this student fails, I have no way to help her perform better in the future.

**Personal Communication**

The final option for assessing mastery of knowledge is direct personal communication with students; for example, by asking questions and evaluating answers. This is a good match across all grade levels, especially with limited amounts of knowledge to be mastered, few students to be assessed, and in contexts in which we need not store records of performance for long periods of time.

The reason I impose these conditions is that this obviously is a time- and labor-intensive assessment method. So if our domain of knowledge to assess is large, we are faced with the need to ask a large number of questions to cover it well. That just doesn’t fit the resource realities in most classrooms. Further, if the number of students to be assessed is large, this option may not allow enough time to sample each student’s achievement representatively. And, if we must store records of performance over an extended period of time, written records will be needed for each student over a broad sample of questions. This, too, eats up a lot of time and energy.
Assessment via personal communication works best in those situations when teachers are checking student mastery of critical content during instruction in order to make quick, ongoing adjustments. Further, sometimes with some students in some contexts, it is the only method that will yield accurate information. For various reasons, some students just cannot or will not participate in the other forms of assessment, such as those who experience debilitating evaluation anxiety, have difficulty reading English, have severe learning or communication disabilities, or simply refuse to “test.”

Assessing Reasoning Proficiency

We constantly use our knowledge to reason and to solve problems. In the previous chapter, we described several valued patterns of reasoning. For example, one is evaluative or critical thinking, the ability to make judgments and defend them through application of standards or criteria. In newspapers, movie or restaurant critics evaluate based on their standards of quality. So, too, can students evaluate the quality of a piece of literature or the strength of a scientific argument by learning to apply certain criteria of quality or standards of excellence. This is evaluative reasoning in action.

Another commonly valued pattern is analytical reasoning, the ability to break things down into component parts and to see how the parts work together. Yet another pattern involves using knowledge to compare and contrast things, to infer similarities and differences.

How does one assess these kinds of reasoning targets in the classroom? Our four methodological choices all provide excellent options when we possess both a clear vision of what we wish to assess and sufficient craft knowledge of the assessment methods.

Selected Response

For example, we can use selected response exercises to determine if students can reason well. We can use them to see if students who have read a story can analyze its elements, compare them, or draw inferences or conclusions. Consider the following examples of questions from a reading test:

- **Analytical reasoning**—Which of the following sequences of plot elements properly depicts the order of events in the story we read today? (Offer alternative orderings, only one of which is correct.)
- **Comparative reasoning**—What is one essential difference between the story we read today and the one we read yesterday? (Offer alternative differences, only one of which is correct.)
- **Drawing conclusions**—If you had to choose a theme from among those listed for the story we read today, which would be best? (Offer response options, one of which is best.)

Assuming that these are novel questions posed immediately after reading the story, so students had no opportunity to memorize the answers, they ask students to dip into their knowledge base (about the story) and use it to reason. Students who
see themselves becoming increasingly proficient at responding to questions like these become increasingly confident readers. This can be powerful.

I continue to be surprised by how many educators believe that selected response exercises can test only recall of content knowledge. While multiple-choice formats certainly can do this very well, they also can tap important reasoning proficiencies.

There are limits, however. *Evaluative reasoning*—the ability to express and defend a judgment, opinion, or point of view—cannot be tested using multiple-choice or true/false items because this kind of reasoning requires at least a presentation of the defense. Answers are not merely right or wrong—they vary in quality. Essay, performance assessment, or personal communication are needed to present that defense.

In a similar sense, problems that are multifaceted and complex, involving several steps, the application of several different patterns of reasoning, and/or several problem solvers working together, as real-world problems often do, demand more complex assessment methods.

But, nevertheless, for some relatively simple, straightforward patterns of reasoning, such as analysis, comparison, classification, and the like, selected response can work. Be advised also that we can provide students with sharply focused practice in mastering valued patterns of reasoning by having them practice writing sample test items that require them to properly reason out the answer. This is assessment FOR learning in action.

**Essay**

This represents an excellent way to assess student reasoning and problem solving. Student writing provides an ideal window into student thinking. Teachers can devise highly challenging exercises that ask students to analyze, compare, draw complex inferences, evaluate, or use some combination of these proficiencies, depicting their reasoning in written form.

Of course, the key to evaluating the quality of student responses to such exercises is for the assessor to understand the pattern of reasoning required and be able to detect its presence in student writing. This calls for exercises that really do ask students to reason through an issue or figure something out, not just regurgitate something that they learned earlier. And these exercises must be accompanied by clear and appropriate scoring criteria that reflect sound reasoning, not just content mastery.

Very often, we can help our students become confident masters of various patterns of reasoning by providing them with the opportunity to assess their own reasoning and problem solving during learning. In this assessment FOR learning context, we might, for example, engage them as partners in creating student-friendly versions of scoring criteria for evaluating the quality of their analytical reasoning.

**Performance Assessment**

Once again, here we have an excellent option that is applicable across all grade levels. We can watch students in the act of problem solving in a science lab, for example, and draw inferences about their proficiency. To the extent that they carry
out proper procedures or find solutions when stymied, they reveal their ability to carry out a pattern of reasoning. When we watch students work with math manipulatives to demonstrate a problem solution or figure out how to manipulate computer software to accomplish something that they haven’t done before, we can literally see their reasoning unfolding in their actions.

However, again, drawing conclusions about reasoning proficiency on the basis of the quality of student products can be risky. If performance is weak, did the student fail to perform because of a lack of basic knowledge, failure to reason productively, or lack of motivation? As previously stated, without followup assessment by other means, we just don’t know. If we don’t follow up with supplemental assessment and thereby infer the wrong cause of failure, at the very least our remedy is likely to be inefficient. We may waste valuable time reteaching material already mastered or teaching reasoning skills already developed.

**Personal Communication**

One of the strongest matches between target and assessment method in Table 4.1 is the use of personal communication to evaluate student reasoning. Teachers can do any or all of the following:

- Ask questions that probe the clarity, accuracy, relevance, depth, and breadth of reasoning.
- Have students ask each other questions and listen for evidence of sound reasoning.
- Have students reason out loud, describing their thinking as they confront a problem.
- Have students recount their reasoning processes.
- Ask students to evaluate each other’s reasoning.
- Simply listen attentively during class discussions for evidence of sound, appropriate reasoning.

Just talking informally with students can reveal so much, when we know what we’re looking for! However, with this method, it will always take time to carry out the assessment and to keep accurate records of results.

**Assessing Mastery of Performance Skills**

When our assessment goal is to find out if students can demonstrate performance skills, such as play a role in a dramatic performance, fluently speak in a second language, effectively give a formal speech, or interact with classmates in socially acceptable ways, then there is only one way to assess. We must observe them while they are exhibiting the desired behaviors and make judgments as to their quality. This calls for performance assessment. There is no other choice. Each of the other options falls short for this kind of target.
But sometimes limited resources make it impossible to assess the actual skill. At those times, we may need to go for second best and come as close to the real target as we can. We have several options when we need to trade high fidelity for greater efficiency in skills assessment. We can use selected response test items to determine whether students can recognize high-level achievement. For example, given a number of performance demonstrations (on video, perhaps), can respondents identify the best? Or, we may use a multiple-choice format to see if students know the proper sequence of activities to carry out when that is relevant to the outcome. Given several descriptions of a procedure, can respondents identify the correct one? We can also use this method to ask if students have mastered the vocabulary needed to communicate about desired skills.

Realize, however, that such tests assess only prerequisite knowledge underpinning effective performance—important building blocks to competence, to be sure. But they will not assess examinees' actual levels of skill in performing. With this same limitation, we could have students write essays about the criteria they might use to evaluate performance in a vocal music competition, knowledge that might well represent an important foundation for performing well in such a competition. But, of course, this will fall short of a real test of performance. Only performance assessment will suffice.

Finally, personal communication represents an excellent means of skills assessment when the skills in question have to do with oral communication proficiency, such as speaking a foreign language. For such an outcome, this is the highest-fidelity assessment option. For other kinds of performance skills, however, personal communication falls short of providing direct data on students' abilities.

Assessing Proficiency at Creating Products

The same limitations discussed for performance skills assessment apply here. If our assessment goal is to determine whether students can create a certain kind of achievement-related product, there is no other way to assess than to have them actually create one. In fact, performance assessment represents the only means of direct assessment. The best test of the ability to throw a ceramic pot is the quality of the finished pot. The best test of the ability to set up a scientific apparatus is the completed arrangement. The best test of the ability to write a term paper is the finished paper.

Again, we could use selected response assessment to see if students can pick out a quality product from among several choices. Or, we could test knowledge of a quality product's key attributes. But these are limited substitutes for assessment that actually asks students to create the product.

It is also possible to have students answer questions, write brief essays, or just discuss informally the key attributes of a carefully crafted object, such as a cabinet in shop class (personal communication). In this way, we can be sure they start with the key understandings they need—a necessary, but not sufficient, condition for success. Then students won’t waste valuable time working on projects they are not prepared to succeed on.
But ultimately the real issue is whether students can create a carefully crafted cabinet. When that is the question, product-based performance assessment is the method of choice.

When we wish to assess writing proficiency by judging the quality of students' written products, we have two methodological choices. One is to have them write brief essays. The other is to have them produce much longer performance assessment products, such as term or research papers. Both are acceptable when accompanied by high-quality scoring criteria reflecting attributes of good writing.

Assessing Dispositions

Let’s take a minute to review some of the student characteristics that fall under this heading. Affective dimensions of individuals that might be the object of classroom assessment include attitudes, values, interests, self-concept, and motivation. Remember, as stated earlier, the focus of assessment in this case is to determine the direction and intensity of student feelings about different school-related issues. When it comes to dispositions, we typically seek strong positive affect: positive attitudes about school, subjects, classmates, and so on; strong values about hard work; a strong positive academic self-concept; and strong positive motivation or seriousness of purpose. But negative attitudes—about drugs, for example—are important, too.

The key to success in assessing these things, as usual, is a clear and appropriate definition of the characteristic to be assessed. Given a clear understanding, could we translate such targets into selected response questions? You bet! But our collection of such items won't be a test per se, it would more properly be considered a questionnaire.

Selected Response Questionnaire

Many selected response formats are very useful for such highly structured questionnaires. For instance, we could offer students statements and ask if they agree, disagree, or are undecided. Or, we could ask them to select from among a list of adjectives those that most accurately apply to themselves or to some other object. This assessment realm is rich with useful options.

Written Response Questionnaire

Essay questions are another viable option for tapping dispositions. We can write open-ended questionnaire items that ask students to describe both the direction and intensity of their feelings about any particular object. After more than 30 years of in-school and in-classroom research on classroom assessment practices, one of the most startling realizations for me has been how rarely teachers use questionnaires to gather affective information from their students, information that could make everyone’s job much easier. In fact, the act of seeking student opinion can yield its own very positive impact on student dispositions. They may be honored to know that you care what they feel about aspects of the classroom.
Chapter 4  Designing Quality Assessments

Performance Observations
The match between dispositions and performance assessment is a bit more complex, however, because of the nature of the inferences we must draw. In this case, I urge caution. It certainly is possible to look at samples of student performance or at student-created products and draw conclusions about attitudes, values, and motivational dispositions with respect to that particular project. If students demonstrate high levels of achievement, their attitude was probably strongly positive, they probably valued the project and their work, and they probably were disposed to work hard to perform well. Just remember, such inferences on your part can be wrong. There is also some chance that such students are just coping well with a frustrating academic challenge, are angry about the project, in fact, and just don’t want you to know it.

Care must be exercised at the low end of the performance continuum, too. When performance is poor, there are many possible explanations, only one of which is a poor attitude and lack of motivation. Only additional followup assessment will reveal the real reason for failure to perform.

Personal Communication
One excellent way to conduct such a follow up is direct personal communication with students. In the right atmosphere, students will talk openly and honestly about the strength and direction of their dispositions. The keys to success, of course, are to be able to establish that open, trusting environment and to know what kinds of questions to ask to tap important affect.

With this brief overview of the match between dispositions and our assessment options, we leave this topic until Chapter 9, which is devoted to assessing student dispositions.

Time for Reflection
To review, please create your own version of the matrix crossing achievement targets with assessment methods (Figure 4.2). Fill in the cells that describe the matches and mismatches in your own words. These connections are critical to your success as a classroom assessor.

The Other Key Design Features of Quality Assessments

We have established the first of our four design features to be considered in creating a quality classroom assessment that meets our goal of gathering valid and reliable evidence of student achievement: We must be sure we select an assessment method capable of reflecting the valued achievement. Different assessment contexts (purposes and targets) afford us opportunities to use different assessment methods. Valid assessments rely on proper methods—methods capable of tapping the learning in question.
As stated earlier, quality assessments also display three design features beyond proper method: they are built of quality ingredients; they provide valid samples of level of achievement mastery; and they are constructed in ways that minimize bias and distortion that can result in mismeasuring achievement. We introduce these three features in the following subsections and will discuss them further in Part II as they relate to each assessment method.

**Design Feature 2: Build the Assessment of Quality Ingredients**

Once we select an assessment method, we need to use that method in a manner that will ensure us an accurate portrait of achievement. To begin with, this means we must build the assessment of high-quality ingredients. If it is to be a multiple-choice test, we must build it of good multiple-choice test items, not poor-quality items. The difference is critical to assessment quality and we will consider it in depth in Chapter 5. A sound performance assessment requires good performance tasks and scoring guidelines, not poor-quality ones. You have to know the difference. Each method brings with it a set of specific procedural guidelines for creating high-quality ingredients. We will explore these in the chapters of Part II.

**Design Feature 3: Sample Achievement Appropriately**

Not only must we create quality ingredients, but we must assemble enough of the right kinds of exercises (test items, for example) to sample student performance in a representative manner. We want to gather enough evidence to lead us to a confident conclusion without wasting time gathering more than we need. Any assessment represents a subset of all the exercises we could have posed if the assessment were infinitely long. A sound assessment relies on a sample that is systematically representative of all the possibilities. We use performance on the sample to infer or generalize about how much of the target sampled each student has mastered.

Our goal in assessment design is to use the most powerful assessment option we can. Power derives from the accuracy and efficiency with which a method can represent our valued achievement target. We always want the highest-resolution picture of that target we can get using the smallest possible sample of student performance; maximum information for minimum cost. I will provide specific guidelines for how to do this for each method in the chapters of Part II.

**Design Feature 4: Minimize Bias**

Further, we must understand that, even if we select a proper method, build it of quality items, and sample appropriately, each method still carries with it a list of things that can go wrong that can mislead us about student achievement. As it turns out, distortions in assessment results can creep in from such sources as the scoring process, the emotional state of the student, and the test administration environment.
Part of assessment literacy is understanding those potential sources of bias and knowing what to do by way of assessment development and implementation to prevent such problems. Again, the chapters of Part II will provide instruction on how to anticipate and avoid bias so as to maximize assessment accuracy.

Assessment FOR Learning

However, before we turn to our discussion of the various assessment methods I would remind you that the basic premise of this book is that we can involve students actively in classroom assessment and thus derive great motivational and learning benefits for them. Here are more suggestions for how you can assess FOR learning:

- Develop a pretest version of a multiple-choice final exam, have students take it early in the unit of study, and help them analyze their results item by item for its knowledge or reasoning attributes, so they can see their own strengths and weaknesses from the beginning of the learning. This begins to focus their attention of the keys to their own ultimate success.
- During a particular unit of study (that is, while students are still learning) have students work in teams to draft practice essay exercises. Then have teams trade exercises and develop scoring criteria for each other's exercises. This can go even further, if you want. Have students try to respond to each other's exercises and practice applying the scoring criteria. Then engage everyone in a discussion of the experience.
- Provide students with a sample of your writing—a piece of work that you have not finished yet—so they can give you some feedback. Then revise your work based on their feedback and show them their impact on the quality of your work.
- Have your students create their own performance assessment exercises that tap performance skill and product targets that you establish as important for them to master. In addition, have them develop scoring guides for evaluating their own performance. Use these for practice before the final assessment.

Sprinkled throughout the remaining chapters are suggestions of ways to involve students to unlock for them the secrets of their own academic success. All four assessment formats welcome student involvement during learning.

Summary: A Vision of Excellence in Classroom Assessment

In this chapter, we made some crucial connections: we connected the why, what, and bow of assessment into one unified picture. Sound classroom assessments arise from a clear sense of purpose, clear and appropriate targets, and reliance on a proper assessment method. A proper method is one that provides the most direct view of student performance, permitting
the strongest inferences from the assessment results to the actual status of the achievement target. Such assessments also are built of quality ingredients, provide valid samples of level of achievement mastery, and are constructed in ways that minimize bias.

We have established four assessment methods: selected response tests, essay exercises, performance assessments, and direct personal communication with students. We discussed how we might use them selectively to tap student achievement on a range of kinds of achievement. Given the range of our valued achievement on a range of kinds of assessment, we might use them selectively to tap student achievement on a range of kinds of achievement. Given the range of our valued achievement targets, we will need to apply all of the assessment tools we have at our disposal—no single method can serve all of our assessment needs at all grade levels. We must learn to use all available methods. If we do, the results will be better information about student success gathered in less time.

Therefore, selecting a proper assessment method becomes our third criterion, and thus the third entry in our set of comprehensive rubrics for evaluating classroom assessment quality (Figure 4.3; the complete set appears in Appendix B). Assessments are ready to go when the method matches the target. They still need work when there is an obvious mismatch.

In the chapters that follow, we provide instruction and practice on how to use each method productively by sampling student achievement properly with high-quality exercises in ways that minimize bias. We also will explore using student-involved assessment as a motivational and teaching tool.

<table>
<thead>
<tr>
<th>Still Needs Work</th>
<th>Well on Its Way</th>
<th>Ready to Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>The method selected is not capable of accurately reflecting the target.</td>
<td>(There is no half way for this one—it either matches or does not.)</td>
<td>A method has been selected that fits the desired achievement target.</td>
</tr>
<tr>
<td><strong>Note:</strong> You will learn much more about the following quality criteria in the chapters of Part II:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The test items, essay exercises, or performance tasks and scoring guides are of poor quality.</td>
<td>Some of the ingredients are of good quality, but others need work.</td>
<td>All items, exercise, tasks, and scoring procedures are of very high quality.</td>
</tr>
<tr>
<td>The sample of items or exercises include clearly do not cover (represent) the domain to be assessed.</td>
<td>The sample is adequate, but could be improved in size and coverage.</td>
<td>The sample collected will lead to confident conclusions about student mastery of the domain.</td>
</tr>
<tr>
<td>Relevant sources of bias remain and will distort results.</td>
<td>Some sources of bias have been eliminated, but some remain.</td>
<td>All relevant sources of bias have been eliminated.</td>
</tr>
</tbody>
</table>

**Figure 4.3**
Guide for evaluating assessments for assessment design
Final Chapter Reflection

1. What are the three most important new insights to come to you as a result of your study of this chapter?
2. Which of your previous questions about assessment can you now answer based on your study of this chapter?
3. What new questions have come to mind as a result of your study of this chapter that you hope to have answered as your study continues?

Practice with Chapter 4 Ideas

1. If your supervisor’s objective is to evaluate your teaching proficiency, what assessment method(s) should be employed and why? (Be careful here! Pause to reflect on all of the active ingredients of good teaching before answering.)
2. When you took your driver’s test, what achievement targets were covered and why? What assessment methods did they use and why?
3. In the vignette presented in Chapter 1, what assessment method did Ms. Weathersby use to determine Emily’s writing proficiency and that of her classmates and why?
4. Draw a blank version of Figure 4.2. In each cell, insert an example of an achievement target from a subject you are familiar with that might be assessed with that method.

Go to our Companion Website at www.prenhall.com/stiggins to:

- assess and apply your understanding of chapter content with multiple-choice questions and essay questions,
- practice your skills with student activities, and
- broaden your knowledge of related issues with Web links to topics on the Internet.