

From CLASSROOM ASSESSMENT FOR  
STUDENT LEARNING: DOING IT RIGHT-USING IT WELL by Richard Stiggins, Judy Arter, Jan  
Chappuis, and Steve Chappuis, Copyright © 2004 by Assessment Training Institute  
(www.assessmentinst.com). Reprinted by permission of the Assessment Training Institute.

PART

2

Assessment Methods



## CHAPTER

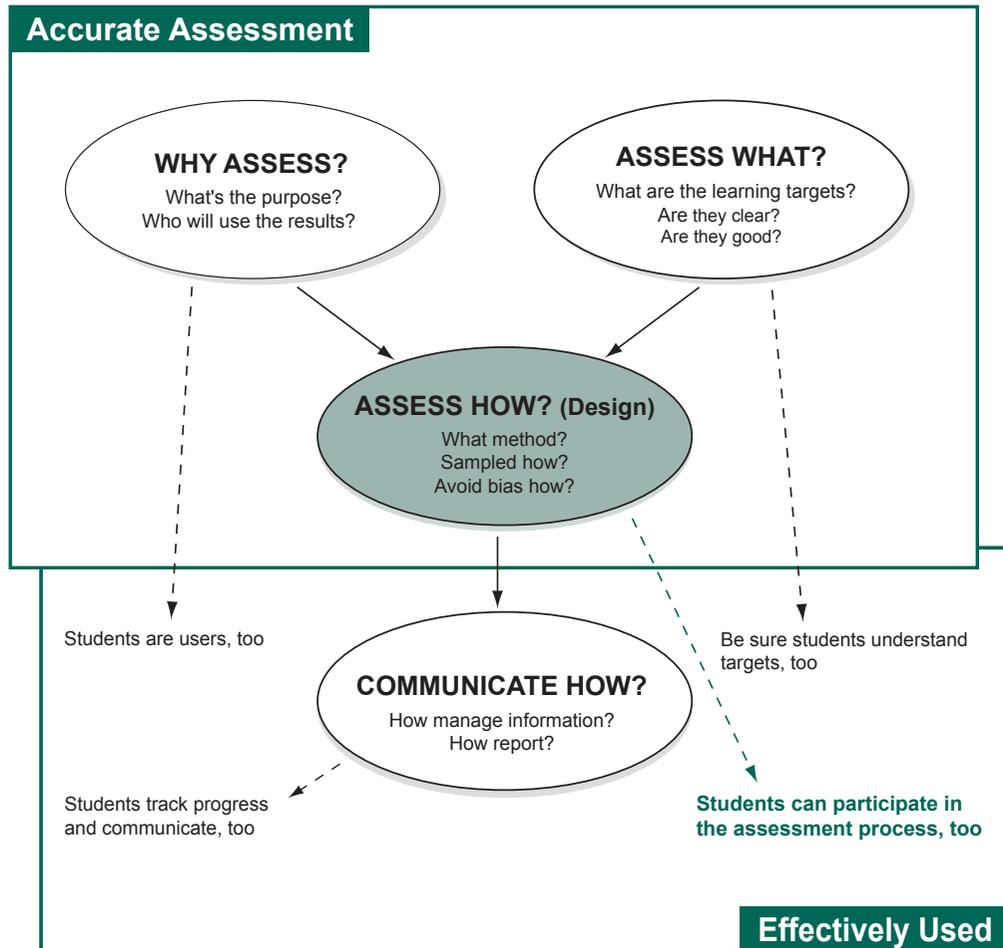
# 4

## Assess How? Designing Assessments to Do What *You* Want

**[ S ]** o far, we have examined two keys to assessment quality. The first key is to know at the outset how we intend to use assessment results. Sometimes we can use them to promote learning (assessment *for* learning) and other times to check to see if learning has occurred—that is, for purposes of accountability (assessment *of* learning). As the second key to quality, we have established that assessments must be designed to reflect the variety of achievement targets that underpin standards: mastery of content knowledge, the ability to use knowledge to reason, demonstration of performance skills and product development capabilities. Now we consider the third key to classroom assessment quality—how to design assessments that cover our targets and serve our purposes (the shaded portion of Figure 4.1).

In this chapter we describe four assessment methods representing the range of assessment options, explain how to choose which method to use for any given learning target, and outline the steps in assessment planning and development. We treat each of the four assessment methods in depth in Chapters 5 through 8; here we offer an overview with an emphasis on selecting the proper method and on thoughtful assessment planning.

Figure 4.1 Keys to Quality Classroom Assessment



## Assessment Methods—A Menu of Options

Throughout our school careers, both as students and as teachers, we have encountered thousands of different assessments. Although the variations are endless, all of the assessments we have experienced and give today fall into one of four basic categories of methods:

1. Selected response and short answer
2. Extended written response

3. Performance assessment
4. Personal communication

All four methods are legitimate options when their use correlates highly with the learning target and the intended use of the information. (Portions of the following discussion are adapted from Stiggins, 2005.)

### Selected Response

Selected response and short answer methods consist of those in which students select the correct or best response from a list provided. Formats include multiple choice, true/false, matching, short answer, and fill-in questions. (Although short answer and fill-in-the-blank do require students to generate an answer, they call for a very brief answer that is counted right or wrong, so we include these options in the selected response category.) For all selected response assessments, students' scores are figured as the number or proportion of questions answered correctly.

### Extended Written Response

Extended written response assessment requires students to construct a written answer in response to a question or task rather than to select one from a list. An *extended* written response is one that is at least several sentences in length. Examples include the following:

- Compare pieces of literature, solutions to environmental problems, or economic events.
- Analyze artwork, forms of government, or solutions to problems.
- Interpret music, scientific information, or polling data.
- Solve a mathematics problem and show and explain all work.
- Describe in detail a scientific, mathematical, or economics process or principle, such as how supply and demand works.

We judge correctness of extended written responses by applying one of two types of predetermined scoring criteria. One type gives points for specific pieces of information that are present. For example, when students in a biology class are asked to describe the Krebs cycle, points might be awarded for noting that the cycle describes the sequence of reac-

tions by which cells generate energy, takes place in the mitochondria, consumes oxygen, produces carbon dioxide and water as waste products, and converts ADP to energy-rich ATP. The second type of criteria can take the form of a rubric, such as a general rubric for making comparisons, which can be applied to any exercise calling for comparison.

Scores therefore also take one of two forms: number or percentage of points attained, or rubric scores.

## Performance Assessment

Performance assessment is assessment based on observation and judgment; we look at a performance or product and make a judgment as to its quality. Examples include the following:

- Complex performances such as playing a musical instrument, carrying out the steps in a scientific experiment, speaking a foreign language, reading aloud with fluency, repairing an engine, or working productively in a group. In these cases it is the doing—the process—that is important.
- Creating complex products such as a term paper, a lab report, or a work of art. In these cases what counts is not so much the process of creation (although that may be evaluated, too), but the level of quality of the product itself.

As with extended written response assessments, performance assessments have two parts: a performance task or exercise and a scoring guide. Again, the scoring guide can award points for specific features of a performance or product that are present, or it can take the form of a rubric, in which levels of quality are described. For example, to assess the ability to do a simple process, such as threading a sewing machine, doing long division, or safely operating a band saw, points might be awarded for each step done in the correct order. Or, for more complex processes or products, you might have a rubric for judging quality that has several dimensions, such as ideas, organization, voice, word choice, sentence fluency and conventions in writing, or content, organization, presentation, and use of language in an oral presentation. Again, scores could be reported in number or percent of points earned, or in terms of a rubric score.

## Personal Communication

Gathering information about students through personal communication is just what it sounds like—we find out what students have learned through interacting with them. Examples include the following:

- Looking at and responding to students' comments in journals and logs
- Asking questions during instruction
- Interviewing students in conferences
- Listening to students as they participate in class
- Giving examinations orally

We usually think of this as informal, rather than formal assessment (in which results are recorded for later use). Often it is. However, as long as the learning target and criteria for judging response quality are clear, information gathered via personal communication can be used to provide descriptive feedback to students, for instructional planning, and for student self-reflection and goal setting. If planned well and recorded systematically, information from personal communication can be used as the basis for assessments *of* learning.

Student responses are evaluated in one of two ways. Sometimes the questions we ask require students to provide a simple, short answer, and all we're looking for is whether the answer is correct or incorrect. This is parallel to scoring for written selected response questions. Questions during instruction usually call for these short answer oral responses.

Other times, student oral responses are longer and more complex, parallel to extended written response questions. Just as with extended written response, we evaluate the quality of oral responses using a rubric or scoring guide. Longer, more complicated responses would occur, for example, during oral examination or oral presentations.

## Misconceptions About Assessment Methods

Much “lore” exists about assessment methods, and we address a few of the most common misconceptions here.

1. *Shouldn't we only be using “authentic” assessments—performance assessments—to judge student progress?* None of these methods is inherently superior to any other, and all are viable if used well. Good assessment means clearly knowing what it is you want to assess and then choosing the best method to get the job done, which, as we will show, depends on the purpose and the learning targets being assessed.
2. *I can see how to involve students in assessment when using a performance assessment, but how do you do it with other methods? Doesn't student involvement require performance assessment?* Although many of our strongest examples of student involvement in the past have come from performance assessment applications, there is a vast, untapped reservoir of student-involvement practices leading to higher learning within each assessment method. You will find extended examples in Chapters 5 through 8.
3. *What about portfolios? I notice they aren't listed as a method. Where do they fit in?* Portfolios are a wonderful idea and we devote an entire chapter to their use later in the book. However, they are not an assessment method, but a vehicle for collecting evidence of, tracking, and communicating about student learning. Portfolios offer a way to involve students deeply in the overall process—self-assessment, tracking progress, reflecting on work, goal setting, and communicating about learning. In this sense, portfolios play a valuable role in creating assessment *for* learning in the classroom.
4. *What about presentations, group projects, worksheets, observations, exhibitions of mastery, posters, and the other ways that teachers gather information?* All of these artifacts and procedures can be classified within the four basic assessment methods described. Presentations and observations are examples of performance assessment. Exhibitions of mastery and group projects can take the form of extended written response, performance assessment, or personal communication depending on how they are carried out. Worksheets are not a method at all because they can contain various types of questions. (Usually worksheets consist of selected response or extended written response questions.) Likewise, posters can be considered either extended written response or performance assessment depending on the assignment and the learning targets being assessed.

## Target–Method Match

One of the values in classifying assessments according to method is that we can think clearly about how to assess what we are teaching. The heart of accuracy in classroom assessment revolves around matching different kinds of achievement targets, with all the forms and nuances of each, to the appropriate assessment method. This is easily done and can save time in the long run.

To begin thinking about the match between kind of learning target and assessment method, please complete the following two activities. You may want to discuss possible answers with colleagues.



### DEEPEN UNDERSTANDING

#### Activity 4.1 Which Method?

Let's say you need to assess student achievement on each of the following learning targets. Which assessment method—selected response/short answer, extended written response, performance assessment, or personal communication—would you choose? Please jot down your answers and save them for later reference.

1. Ability to write clearly and coherently
2. Group discussion proficiency
3. Reading comprehension
4. Proficiency using specified mathematical procedures
5. Proficiency conducting investigations in science



## DEEPEN UNDERSTANDING

### Activity 4.2 Target–Method Match

For this activity, you will determine which assessment method is the best match for each of the four kinds of learning targets: knowledge, reasoning, skill, and product targets. To do this you will need to read through the following four scenarios and record your answer to each question by marking an “X” in the appropriate box on Figure 4.2 when your answer is “Yes.” (A printable version of the figure is on the accompanying CD in the file, “Target–Method Match Chart.”) You can put an “X” in more than one box. You can use capital “X” to denote really good matches, and a checkmark to denote an acceptable match under certain conditions (or whatever easily distinguishable marks you wish). On a separate sheet of paper, write your justifications for each answer, as requested. To make your choices, think about accuracy and efficiency: which methods will provide the most accurate information with the highest degree of efficiency? If you are working with a learning team, consider discussing your responses as a group.

#### **Scenario 1: Assessing Student Mastery of Content Knowledge**

*Scenario:* You want your students to master specific subject matter knowledge because it represents an important foundation for later work. You plan a series of instructional activities to help your students reach this goal. Now you want to assess to be sure they’ve got it. In this particular case, you want them to know the material outright, not through the use of reference materials.

*Question 1:* Should you assess mastery of this material using selected response or short answer modes of assessment, such as multiple choice, true/false, or matching exercises? Briefly explain your response.

*Question 2:* Should you assess your students’ mastery of this material using an extended written response form of assessment? Defend your answer.

*Question 3:* Should you use a performance assessment to assess students’ mastery of this content knowledge? Defend your answer.

*Question 4:* Do you think the personal oral communication form of assessment—by oral exam, interview, conference, or discussion—could viably assess your students’ mastery of this content knowledge? Why or why not?

## Activity 4.2 (Continued)

### **Scenario 2: Assessing Reasoning Proficiency**

*Scenario:* You are a teacher who has seen to it that your students are able to access important knowledge when required. Now you want to see if they can use that knowledge productively to solve relevant problems. You want to see if they can reason analytically (think about the parts of things) and comparatively (think in terms of similarities and differences), draw inferences, and think critically (take and defend a position on an issue, for example).

*Question 1:* Can you get at these things with selected response or short answer assessments? Why or why not?

*Question 2:* Does extended written response assessment work in contexts where we seek to assess reasoning proficiency? Why or why not?

*Question 3:* Is performance assessment a viable alternative? Why or why not?

*Question 4:* Can we use personal oral communication as an assessment method to probe a student's ability to reason effectively and solve problems? Defend your response.

### **Scenario 3: Assessing Mastery of Skills**

*Scenario:* You teach French and wish to assess your students' skill at communicating in that language in a conversational situation. So the skill of *oral language proficiency* is your target.

*Question 1:* Can you assess oral language proficiency in a conversational context using a selected response or short answer mode of assessment? Defend your answer.

*Question 2:* Can you assess these skills using extended written response assessment? Why or why not?

*Question 3:* Will performance assessment work as a basis for assessing the foreign language speaking proficiency of your students? Why or why not?

*Question 4:* Can you use personal oral communication as a basis for assessing conversational skill in a second language? Defend your response.

*Question 5:* Would your responses also apply to other skills such as operating a sewing machine, dribbling a basketball, or reading aloud fluently?

## Activity 4.2 (Continued)

### **Scenario 4: Assessing the Ability to Create Quality Products**

*Scenario:* You want your students to be able to create quality products—products that meet certain specified standards. They might be samples of writing, term papers, technology products, craft products, artistic creations, or others. Your instruction has centered on helping students learn the differences between products that are of high and low quality. You have provided practice in developing products that meet your standards. Now it is time to assess the students' achievement to see if your instruction was effective.

*Question 1:* Can you assess the ability to create these kinds of products using selected response or short answer modes of assessment? Why or why not?

*Question 2:* Will extended written response assessment work for evaluating this kind of achievement? Explain.

*Question 3:* Can performance assessment provide the evidence of proficiency needed to evaluate this kind of achievement target? Defend your response.

*Question 4:* Is personal oral communication a viable way to assess when creation of a product is the target? Why or why not?

### **Checking Your Matches**

Table 4.1 identifies the strong matches between kinds of achievement targets and assessment methods. Please compare the table to the target–method match choices made by your team. Note and discuss discrepancies.

**Figure 4.2 A Plan for Matching Assessment Methods with Achievement Targets**

Target to Be Assessed	Assessment Method			
	Selected Response	Extended Written Response	Performance Assessment	Personal Communication
Knowledge Mastery				
Reasoning Proficiency				
Performance Skills				
Ability to Create Products				

Source: Adapted from *Student-Involved Assessment for Learning*, 4th ed. (p. 65), by R. J. Stiggins, 2005, Upper Saddle River, NJ: Merrill/Prentice Hall. Copyright ©2005 by Pearson Education, Inc. Adapted by permission of Pearson Education, Inc.

## Assessing Knowledge Targets

### *Selected Response*

There is usually no argument about this match. Selected response options do a good job at assessing mastery of discrete elements of knowledge, such as important history facts, spelling words, foreign language vocabulary, and parts of plants. These assessments are efficient in that we can administer large numbers of questions per unit of testing time and so can cover a lot of material relatively quickly. Thus, it is easy to obtain a good sample of student knowledge so that we may infer level of overall knowledge acquisition from the sample on the test.

**Table 4.1 Links Among Achievement Targets and Assessment Methods**

Target to Be Assessed	Assessment Method			
	Selected response	Extended Written Response	Performance Assessment	Personal Communication
<b>Knowledge Mastery</b>	Good match for assessing mastery of elements of knowledge.	Good match for tapping understanding of relationships among elements of knowledge.	Not a good match—too time consuming to cover everything.	Can ask questions, evaluate answers and infer mastery—but a time-consuming option.
<b>Reasoning Proficiency</b>	Good match only for assessing understanding of some patterns of reasoning.	Written descriptions of complex problem solutions can provide a window into reasoning proficiency.	Can watch students solve some problems and infer reasoning proficiency.	Can ask student to “think aloud” or can ask followup questions to probe reasoning.
<b>Skills</b>	Not a good match. Can assess mastery of the knowledge prerequisites to skillful performance, but cannot rely on these to tap the skill itself.		Good match. Can observe and evaluate skills as they are being performed.	Strong match when skill is oral communication proficiency; not a good match otherwise.
<b>Ability to Create Products</b>	Not a good match. Can assess mastery of knowledge prerequisite to the ability to create quality products, but cannot use to assess the quality of products themselves.	Strong match when the product is written. Not a good match when the product is not written.	Good match. Can assess the attributes of the product itself.	Not a good match.

Source: Adapted from *Student-Involved Assessment for Learning*, 4th ed. (p. 69), by R. J. Stiggins, 2005, Upper Saddle River, NJ: Merrill/Prentice Hall. Copyright ©2005 by Pearson Education, Inc. Adapted by permission of Pearson Education, Inc.

### *Extended Written Response*

Extended written response is useful for assessing blocks of knowledge rather than pieces of information detached from one another—causes of environmental disasters, the carbon cycle in the atmosphere, how one mathematical formula can be derived from another, or the concept of checks and balances in government. Extended written response assessments are not as efficient as selected response assessments because responding to each takes longer, but they can get at deeper levels of knowledge.

### *Performance Assessment*

Performance assessment is usually not a good choice for assessing knowledge targets, for three reasons.

We'll illustrate the first reason with a brief example. Let's say we ask a student to complete a rather complex performance, such as writing and executing a computer program, for the purpose of determining if she has the prerequisite knowledge. If the student successfully executes the program, then we know that she possesses the prerequisite knowledge. The problem comes in when the program does not run successfully. Was it due to lack of knowledge of the programming language, due to the inability to use knowledge to create a program that does what it is intended to do, or merely due to the inability to manipulate the keyboard or to proofread? We can't know the reason for failure unless we follow up the performance assessment with one of the other assessment methods. We must ask some short answer or extended response questions to find out if the prerequisite knowledge was there to start with. But, if our initial objective was to assess mastery of specific knowledge, why go through the extra work? To save time and increase accuracy, we recommend using selected response, short answer, and extended written response assessments to evaluate knowledge targets.

The second reason this is not a good match is because it is inefficient to assess all content knowledge with a performance assessment. A single performance task does require some subset of knowledge, and you can assess its presence with a particular performance task, but how many performance tasks would you have to create to cover all the knowledge you want students to acquire? For example, how many performance assessments would it take to determine if students can spell all the words you want them to spell? Or, how many performance assessments would it take to determine if students can perform all the mathematical operations they have been taught in a semester? Again, we recommend assessing knowledge with a simpler method and reserving performance assessment for those learning targets that really require it.

The third reason that performance assessments are usually not a good match for knowledge learning targets has again to do with practicality. It just isn't practical (or safe) to conduct some performance assessments. For example, let's say that you want to know if students can read schedules, such as bus schedules. It would be most "authentic" to ask students to get around town on the bus, but it would be highly inefficient and perhaps dangerous. Asking students to answer multiple-choice or short answer questions requiring understanding of a bus schedule would be a good compromise for getting the information needed.

### ***Personal Communication***

This is a good match with knowledge targets for most students at all grade levels, but tends to be inefficient if a lot of knowledge is to be assessed for lots of students. Personal communication works best for real-time sampling of student understanding during instruction. Also, for some students, such as those with special needs, English language learners, or younger students, it is the best way to gather accurate information.

## **Assessing Reasoning Proficiency**

### ***Selected Response***

A common misunderstanding is that selected response questions can't assess reasoning proficiency. Although not a good choice for some patterns of reasoning, other patterns of reasoning can be assessed in selected response format. For example:

- Which of the following statements best describes how dogs in real life are different from the dog in the story? (Comparative reasoning)
- What generalization can you make from this selection about how these plants lure their prey? (Inference—generalizing)
- Which answer best explains the author's purpose in writing this story? (Inference—determining author's purpose)
- Choose the sentence that best tells what the story is about. (Inference—identifying main idea)

There are limits to this format when assessing reasoning. If you want to assess how well students can choose from their store of reasoning proficiencies to solve a problem, solve a problem requiring several steps, explain their choice or reasoning process, or defend an opinion, you must use another assessment method. For example, you might ask students

to solve the following problem in mathematics: “Estimate the number of hours of TV advertising the typical U.S. fifth grader watches in a year. Describe your procedure for determining your answer.” This is an extended response question. If the learning target you want to assess is student reasoning, a single number as the right answer is not the focus of the assessment—the process itself is.

### *Extended Written Response*

Extended written response is a good choice for assessing reasoning targets. Students can be encouraged to write to examine their own thinking. The trick here is to write good questions, ones that require students to analyze, compare, contrast, synthesize, draw inferences, and evaluate *novel* information. For example, if you want students to be able to determine the type of government present in a country (such as democracy, theocracy, dictatorship, or monarchy), you could teach the characteristics of each and practice identifying them. Then, on the assessment, you would give the characteristics of a particular government in a country they had not studied and ask students to identify its type and to explain their reasons.

Remember that to assess reasoning, the question has to be novel. If students worked on the answer to the question during instruction, then the answer is a piece of remembered knowledge, which does not require reasoning. For example, consider the following science standard: “Recognize and analyze alternative explanations and models” (National Research Council, 1996, p. 175). The intent of this standard is that students will be able to use scientific criteria to determine the most plausible explanation or model. To assess this aspect of scientific inquiry, students must be asked to evaluate an explanation or model different than the ones the teacher has used for demonstration and also different from the ones the students have practiced with.

### *Performance Assessment*

This is a partial match for assessing reasoning. For example, we can observe students carrying out science laboratory procedures and draw conclusions about their reasoning based on our observations. But, there’s a hitch that keeps performance assessment from being a great match with reasoning targets: we need to make an inference from what we observe. If students do well on a performance task requiring specific patterns of reasoning, we can assume that reasoning is sound. However, if they don’t do well, it could be due to lack of prerequisite knowledge, lack of motivation, or to imprecise reasoning. Without engaging in additional time-consuming assessment, we may not be able to judge level of achievement on reasoning targets.

### ***Personal Communication***

For gathering accurate information, personal communication is a strong match to reasoning targets. Teachers can ask students questions to probe more deeply into a response. Or, students can demonstrate their solution to a problem, explaining their reasoning out loud as they go.

The drawbacks with using personal communication to assess reasoning proficiency are, as always, the amount of time it takes and the record-keeping challenge it poses.

### **Assessing Performance Skills Targets**

There is really only one assessment method that adequately covers performance skills targets, and that is performance assessment. We can use other assessment methods to determine if students possess the knowledge required to perform skillfully, but the only way to determine whether students can actually perform skillfully is to watch them do it and then judge their level of achievement. For example, we can ask students to answer selected response or oral questions about how to conduct themselves during a job interview, but the only way to determine how well they can do it is to watch them during a simulated job interview.

Performance assessment overlaps with personal communication when the performance skills in question fall into the category of oral proficiency, such as speaking a foreign language or giving an oral presentation.

### **Assessing Proficiency in Creating Products**

As with performance skills, the only option for determining whether students can create a certain kind of product is performance assessment: have them create the product or performance and then judge its quality. Once again, we can assess the knowledge required for creating a quality product with a less time-consuming method, but the only way to determine students' levels of proficiency in creating the product is to have them create it.

Performance assessment overlaps with extended written response when the product in question requires writing, such as writing a business letter, lab report, research report, or health and fitness plan.

## Thinking About Target–Method Match

Take a look back to Activity 4.1. See if you'd like to amend any of your original decisions and explain why. Then read our suggested answers here and see if you agree. You may want to discuss your responses and thoughts with your learning team.

1. *Writing proficiency*

The act of writing creates a product that is then assessed with a scoring guide describing components of quality, such as: ideas, organization, voice, word choice, sentence fluency, and conventions. We might also look at the process of writing to observe thinking strategies, planning strategies, revision techniques, or editing proficiency, which would also require a performance assessment.

2. *Group discussion proficiency*

Group discussion is a performance skill—to make a judgment about a student's level of proficiency, we have to observe it. Therefore, a performance assessment is the best option.

3. *Reading comprehension proficiency*

This situation is more complicated. Reading comprehension is an act of reasoning. While students must have decoding proficiency, knowledge of vocabulary, and so forth to comprehend, these are prerequisites that would be assessed prior to determining level of comprehension. Since comprehension is reasoning, we must first define the kind or kinds of reasoning we want to assess. Then we could choose to assess it with selected response (e.g., which of following is most likely to happen next?), extended written response (e.g., list the main idea and two supporting details, or summarize the passage), or personal communication (e.g., orally retell the story, or orally answer comprehension questions).

You might be thinking, “Couldn't we also use performance assessment, like having the students create a poster depicting the main sequence of events in a story, or create a picture that depicts the character of the protagonist?” If what you are assessing is comprehension, the poster or artwork is just the context in which we are getting the extended written response so that we can judge comprehension. We don't want to use information about the quality of the poster or picture as a product itself as evidence of reading comprehension.

4. *Proficiency using specified mathematical procedures*

“Using mathematical procedures” might imply knowledge—ability to carry out the steps in a procedure—or it might imply reasoning—understanding when to use a mathematical procedure. You could use selected response, extended written response, or personal communication to assess either a knowledge or reasoning interpretation.

5. *Proficiency conducting labs in science*

Proficiency conducting labs in science is a performance skill—skillfully using equipment—therefore it requires a performance assessment—watching students use the equipment.



**TRY THIS**

**Activity 4.3 Analyze Samples for Target–Method Match**

In the CD file, “Assessments to Evaluate,” examine the samples entitled “Mathematics Assessment and Instructional Guide,” “Fish Tank,” and “Emerson Essay Test,” and decide the extent to which each has used an appropriate assessment method for the learning targets being assessed. To decide, you’ll first need to identify the kind of learning target and the assessment method. Then refer to Table 4.1 to determine the quality of the match.

Now, look at some of the assessments you have used in the past and decide the extent to which each has used an appropriate assessment method for the learning targets being assessed. To decide, you’ll first need to identify the kind of learning target and the method of assessment. Then refer to Table 4.1 to determine the quality of the match. The learning targets might not be clear enough to decide whether the method(s) are appropriate. If so, make a note to that effect and decide what to do about the situation if you want to use the assessment again.

## Assessment Development Cycle

All assessments, regardless of method selected, go through the same five stages of development: planning, developing, critiquing, administering, and revising (Figure 4.3). So far in the text, we have discussed three of the four steps in the first stage: determine the intended uses of an assessment, identify the learning targets to be assessed, and select

**Figure 4.3 Stages in Assessment Planning**

1. **Plan:** Assess why? Assess what? Assess how? How Important?
2. **Develop:** Determine the sample. Select, create, or modify test items or tasks and scoring mechanisms.
3. **Critique:** Evaluate for quality.
4. **Administer:** Administer the test or assessment.
5. **Revise:** Evaluate test quality based on results and revise as needed.

the proper assessment method. The fourth step, which we address later in this section, is to determine the relative importance of each learning target so that we sample each adequately.

In the second stage we select or create test items or tasks and scoring mechanisms, adhering to the guidelines offered for each method in Chapters 5 through 8.

During the third stage, we check to make sure we have avoided all possible things that might inadvertently cause results to misrepresent student learning, again using information provided for each method in Chapters 5 through 8.

In the fourth stage, we simply administer the assessment to students.

In the fifth and last stage, we note any problems with the questions, tasks, or scoring mechanisms on the assessment and rework them as needed.

The five stages of development we describe here are presented in the context of a teacher-developed assessment for classroom use. However, they also apply to any other type of assessment developed by grade level teams, content area departments, or district subject-area teams for purposes other than individual classroom use. Short-cycle, common, or interim assessments also need to adhere to standards of quality, and the five stages of development should frame that assessment development process, as well. In Chapters 5 through 8 we will describe any variations on the theme applicable for particular assessment methods.

## Stage 1: Plan the Assessment

Creating or selecting a test without having a test plan can result in mismatches between instruction and assessment. The assessment probably will not measure what you intend it to measure, which is known as a *validity* problem. From an assessment quality point of view, this is a bad thing. If you have ever faced an exam yourself that did not match what you thought were the most important aspects of the course you were taking, you know what that feels like from the student's point of view. In the following activity, you will analyze a test you have given to determine its match to the intended learning targets.

When we make a plan for an assessment, whether we intend to create the assessment or just copy it, we are making the advance decisions about validity—what the test will cover and how much weight each learning target will get.



### TRY THIS

#### Activity 4.4 Analyze Your Own Assessment for Clear Targets

Find a selected response test or an extended written response test you have given to students in the past or one you plan to give. Then follow these steps to audit it for clear targets.

1. *Analyze your test item by item.* Identify and write down what learning each item assesses. Describe the learning in whatever terms you want. If two or more items address the same learning, use the same terms to describe that learning. Note the number of test points each item is worth.
2. *Organize the learning targets into a test plan.* Transfer the item information to the chart in Table 4.2 (a printable copy of this table appears on the accompanying CD in the file, “Analyze for Clear Targets”).
3. *Question your test plan.* Is this a representative sample of what you taught and what you expected students to learn?
  - Does the number of points for each learning target represent its relative importance within the whole? If not, which ones are out of balance?
  - Does the number of points for each learning target represent the amount of time you spent on it relative to the whole? If not, which ones are out of balance?
  - Are any important learning targets you taught left out? If so, which one(s)?

**Activity 4.4 (Continued)**

4. *Adjust your test plan.* As needed, adjust the numbers in the “Number of Questions” and/or “Points” column on the table to reflect the amount of time you spent teaching each learning target and each target’s relative importance to the content as a whole.

As needed, add or delete learning targets to reflect what you taught and what you deemed most important to learn and assess.

5. *Draw conclusions about your assessment.* What does the data you wrote into Table 4.2 tell you about the matches among what’s written in your curriculum, what you taught, and what you assessed?

**Table 4.2 Analyze an Assessment for Targets**

Learning Target	Number of Items	Points

### Examples at the Assessment Planning Stage

We'll examine the planning stage by following two examples: a secondary music teacher planning an assessment for his unit on the bluegrass music and a fourth-grade teacher planning an assessment for a unit on the physics of sound. The music teacher's assessment plan is presented in Table 4.3, and the fourth-grade teacher's assessment plan is presented in Table 4.4. As you read these tables, please note that there is no single correct format for test plans; we simply have shown two possible formats.

**Table 4.3** Test Plan—Bluegrass Music

Content	Know	Analyze	Compare	Total
Individual pieces of music	5		5	10
Musical elements/instruments/ expressive devices	5	5		10
Composers/performers	5	5	5	15
TOTAL	15	10	10	35

### Assess Why?

As we saw in Chapter 2, assessment results can be used for many purposes. In each of our two examples, the teachers' primary purposes are twofold: to help students understand how much they have learned, and to add information to the gradebook in preparation for calculating a course grade. Because assessment design is influenced by how we intend to use the results and by whom else will also use them, we answer the question, "Assess why?" first of all.

### Assess What?

Sound assessments arise from clear, specific, and appropriate achievement targets. Beginning with clear targets is important because different targets require different assessment methods and also because the breadth and depth of a learning target will affect how much coverage it will need on the assessment and in instruction. So at this juncture, you will do the following:

1. List the major learning targets you will be teaching.
2. Identify the prerequisite subtargets by unpacking or clarifying the learning targets, as needed.
3. Classify the targets, subtopics, and/or unpacked learning targets, into knowledge, reasoning, performance skills, products, and/or dispositions.
4. Write the unpacked and/or clarified learning targets into the appropriate spaces in the test plan format you select. Blank forms are on the CD in the file, “Test Planning Forms.”

**Table 4.4 Test Plan—The Physics of Sound, for Fourth Graders (Selected Targets)**

Learning Target	Type of Target	Assessment Method	Percent Importance
Acquire vocabulary associated with the physics of sound.	Knowledge	Selected Response	25%
Learn that sound originates from a source that is vibrating and is detected at a receiver such as the human ear	Knowledge	Selected Response	5%
Use knowledge of the physics of sound to solve simple sound challenges	Reasoning	Extended Written Response	20%—Present novel sound challenge; student describes how to solve it
Understand the relationship between the pitch of a sound and the physical properties of the sound source (i.e., length of vibrating object, frequency of vibrations, and tension of vibrating string)	Reasoning	Extended Written Response	10%—Give two novel examples and student compares pitch
Use scientific thinking processes to conduct investigations and build explanations: observing, comparing, and organizing (1) How sound travels through solids, liquids, and air; (2) Methods to amplify sound at the source and at the receiver	Reasoning Skill	Extended Written Response & Performance Assessment	40%—Design an experiment for a given hypothesis; give data/student organizes; set up stations/students conduct an experiment—all novel

Source: From the *FOSSR Physics of Sound Teacher Guide*, © The Regents of the University of California, 2005, developed by Lawrence Hall of Science and published by Delta Education, LLC. Reprinted by permission.

The secondary school music teacher whose test plan is represented in Table 4.3, has planned a 3-week unit of instruction on bluegrass music. He has chosen bluegrass music as the context for the following music standards:

- Classifies selected exemplary works from various historical periods by genre, style, and composer.
- Explains how use of specific musical elements (for example, rhythm, melody, timbre, expressive devices) is characteristic of music from various world cultures.
- Identifies music that represents the history and diverse cultures of our state.
- Identifies important composers and performers who influenced various genres of American music.

Students will need to acquire some knowledge about bluegrass music in three categories—works (famous pieces of music), musical elements (used to give the music the bluegrass feel), and composers/performers. In addition, the teacher will teach students to use the content knowledge in each of these three areas to reason analytically and comparatively. As indicated in the test plan, any single test question either will test knowledge or will be a combination of knowledge and the reasoning that is to be performed using that knowledge.

In the plan for the fourth-grade unit on the physics of sound, the teacher has written selected learning targets down the left-hand column of Table 4.4. The type of learning target is noted in the next column.

These teachers chose content categories based on their state content standards, local curriculum guides, and natural subdivisions of content. They chose reasoning patterns from content standards, local curriculum, and priorities in their teaching.

### **Assess How?**

This is fairly straightforward. Once you have classified learning targets by type it is easy to decide which assessment method to select by referring to the matching guidelines in Table 4.1.

The fourth-grade teacher is emphasizing science process skills as well as knowledge and reasoning so she will be using more than one assessment method. She has chosen the planning format shown in Table 4.4, which allows her to specify how each learning target will be assessed. The music teacher has only knowledge and reasoning learning targets. He has decided that the combination of knowledge and reasoning can be assessed well with a

selected response test. Since he has no need for a test plan to show different assessment methods, he has chosen a test plan format that emphasizes how content knowledge crosses with level of thinking.

### *How Important?*

When we define the relative importance of each of the learning targets listed, we are mapping out how we will *sample* student learning. What will be most important on this assessment? How many points will each item be worth? For the most part, this is the call of the individual teacher, taking into account the following:

- *The breadth and depth of the learning target.* For example, in Table 4.4, the learning target “Learn that sound originates from a source that is vibrating and is detected at a receiver such as the human ear” doesn’t cover as much territory as “Acquire vocabulary associated with the physics of sound,” or “Use scientific thinking processes to conduct investigations and build explanations: observing, comparing, and organizing.” Therefore, assessing “learning where sound originates” will carry less weight on the assessment, as reflected by the percentage of total points, and other targets will carry more weight.

In all cases, the assessment must include enough questions or tasks to provide evidence leading us to a confident conclusion about student achievement, without wasting time gathering too much evidence. The critical question is, How much evidence is enough? How many multiple choice test items, essay exercises, performance tasks? (Each assessment method brings with it a set of rules of evidence for determining how big a sample of student achievement we need. We explain those guidelines in Chapters 5 through 8.)

- *The importance of each learning target.* For example, in Table 4.4, the teacher has determined that the most important learning target focuses on science processes and skills. Scientific information is important, and there is an expectation that students will learn some content information from this unit of study, but process skills are more important in this case. Therefore, science process targets alone will comprise 40 percent of the assessment points and the other four targets combined will total 60 percent.
- *State standards and local curriculum.* For example, the music teacher is guided by the state standard in his emphasis of knowledge and reasoning targets in the unit. Because the state standards emphasize using information to analyze and classify, the teacher has also emphasized it on his test—two-thirds of the points on the test reflect students’ ability to apply knowledge in novel ways.

Although not a hard and fast rule, a good guideline for making decisions regarding percentage of importance for each learning target is that percentage of instructional time and percentage of assessment time should be roughly equal. So, if science processes and skills represent 40 percent of importance, roughly 40 percent of instructional time will be used to teach science processes and skills.

## **Stage 2: Develop the Assessment—Good Exercises and Sound Scoring Procedures**

Having chosen one or more assessment methods, we must adhere to guidelines for developing test questions, extended written response exercises, performance tasks, or questions to elicit evidence of the desired student learning. Further, we need accurate scoring keys for selected response assessments and good scoring guides for extended written response and performance assessment. The development guidelines for each are covered in detail in Chapters 5 through 8.

## **Stage 3: Critique the Assessment**

We've carefully selected and clarified learning targets, determined their relative importance, chosen the best assessment method, and carefully written questions, tasks, and scoring guides adhering to standards of quality. We're finished, right?

Well, no. Regardless of how carefully we plan, things can still go wrong that result in inaccurate estimates of achievement. Witness the pumice that wouldn't float in one state's performance assessment of science process skills, the writing prompt that elicited expository instead of persuasive writing from students, the 10th-grade standardized test administered during rioting by seniors, or asking English language learners to write extended responses to math problems.

A list of problems that can result in inaccurate estimates of student learning is presented in Figure 4.4. Note that some of the problems listed, such as unclear targets, inappropriate assessment method, and improper sampling, would be solved by adhering to the test development process as discussed here. Others are crazy little things that can be hard to anticipate if you haven't experienced them. Problems specific to each method will be discussed more fully in Chapters 5 through 8.

**Figure 4.4 Potential Sources of Bias and Distortion**

- 1. Potential barriers to accurate assessment common to all methods**
  - A. Barriers that can occur within the student**
    - Language barriers
    - Emotional upset
    - Poor health
    - Physical handicap
    - Peer pressure to mislead assessor
    - Lack of motivation at time of assessment
    - Lack of testwiseness (understanding how to take tests)
    - Lack of personal confidence leading to evaluation anxiety
  - B. Barriers that can occur within the assessment context**
    - Noise distractions
    - Poor lighting
    - Discomfort
    - Lack of rapport with assessor
    - Cultural insensitivity in assessor or assessment
    - Lack of proper equipment
  - C. Barriers that arise from the assessment itself (regardless of method)**
    - Directions lacking or vague
    - Poorly worded questions
    - Poor reproduction of test questions
    - Missing information
- 2. Potential barriers to accurate assessment unique to each method**
  - A. Barriers with multiple-choice tests**
    - Lack of reading skills
    - More than one correct response choice
    - Incorrect scoring key
    - Incorrect bubbling on answer sheet
    - Clues to the answer in the item or in other items

**Figure 4.4 (Continued)**

**B. Barriers with extended written response assessments**

- Lack of reading or writing skills
- No scoring criteria
- Inappropriate scoring criteria
- Evaluator untrained in applying scoring criteria
- Biased scoring due to stereotyping of respondent
- Insufficient time or patience to read and score carefully
- Students don't know the criteria by which they'll be judged

**C. Barriers with performance assessment**

- Lack of reading skills
- Inappropriate or nonexistent scoring criteria
- Evaluator untrained in applying scoring criteria
- Bias due to stereotypic thinking
- Insufficient time or patience to observe and score carefully
- Student doesn't feel safe
- Unfocused or unclear tasks
- Tasks that don't elicit the correct performance
- Biased tasks
- Students don't know the criteria by which they'll be judged
- Insufficient sampling

**D. Barriers when using personal communication**

- Sampling enough performance
- Problems with accurate record keeping

Source: Adapted from *Practice with Student-Involved Classroom Assessment* (pp. 194–195), by J. A. Arter & K. U. Busick, 2001, Portland, OR: Assessment Training Institute. Adapted by permission.

### Stage 4: Administer the Assessment

If you have completed the steps in the preceding three stages, administering the test should go smoothly and the test itself will meet the information needs of all parties involved. Unfortunately, administering the test has too often been the starting place and the “cycle” may have looked rather linear: teach, test, report. If we have used textbook tests and relied on the textbook company to take us through stages one through three, we may have high-quality items, but we will not have attended to the need to make the test reflect the balance of importance of what we taught. Because textbook tests are generally designed to meet only the teacher’s need to generate a grade, it may be quite cumbersome to use them to provide students with specific information regarding their strengths and areas of need, standard by standard. We recommend that you never begin the assessment development cycle at Stage 4.

### Stage 5: Revise the Assessment

At this stage, we come full circle to double check that the test did indeed accomplish what we intended. Were we able to use the results for all the decisions we intended to make? Were students able to use the results to keep in touch with the details of their progress? Did a source of bias or distortion creep in? It is almost impossible to eliminate *all* sources of bias and distortion up front. Some only become apparent when you give students the assessment. So,

1. Do the best you can prior to administering the assessment.
2. Watch for possible sources of error during and after the assessment.
3. If something goes wrong, either (1) don’t use the results from the tasks in question, or (2) interpret the results with possible bias in mind.

Consider asking students to help in this. Were parts of the test a surprise to them? Did aspects seem not to match their understanding of what it was most important to learn? Were some questions or tasks confusing—the students knew the material, but didn’t know how to respond? Remember that our goal in the classroom is to get accurate information about student achievement, and if we know the information is not accurate, we have an obligation to discard it and to revise the assessment, if needed, before future use.

## Assessment for Learning Using Assessment Plans

Assessment *for* learning and student involvement activities spin directly off the assessment plan. For example, the plan itself can be shared with students ahead of time to make the learning targets clearer. Students can be involved throughout a unit, by identifying where each day's instruction fits into the plan, or by writing practice test questions periodically for each cell of the plan, as a form of review. More ideas for using test plans as instructional tools will be described in Chapters 5 and 6.

For performance assessment targets, students can be given rubrics for the reasoning and skills outlined in the assessment plan and can practice using the rubrics throughout instruction. For example, in the fourth-grade unit on the physics of sound, the teacher has rubrics for what makes good observation, comparing, and organizing. Students can use these rubrics to guide their activities during the experiments outlined in the unit. Then the same rubrics determine success on a final set of culminating activities. The same set of rubrics could also be used in any other unit emphasizing the same reasoning or skill targets, thus building student understanding and competence over time. These and other procedures for using rubrics as tools for learning are examined in detail in Chapter 7.

## Summary

No single assessment method is superior to any other. Selected response, extended written response, performance assessment, and personal communication are all viable options depending on the learning targets to be assessed, the purpose of the assessment, and special student characteristics such as age, English proficiency, or specific learning disabilities.

All assessment development proceeds through the same five stages: (1) identify the purpose, specify the targets, select appropriate methods, decide on relative importance of the targets and sample well; (2) write the questions using guidelines for quality; (3) eliminate as many potential sources of bias and distortion as possible; (4) administer the assessment; and (5) examine the results for areas needing fine tuning. By doing the work at each stage, we can have confidence that our assessments are yielding accurate results.



## DEEPEN UNDERSTANDING

### **Activity 4.5 Video Discussion of *Evaluating Assessment Quality***

ATI's interactive video, *Evaluating Assessment Quality: Hands-On Practice*, provides review and extra practice on characteristics of quality classroom assessments—clear purposes, clear targets, target–method match, sampling, and potential sources of bias and distortion—using workshop-type activities. We recommend working through the video with your learning team.

### ■ ***Tracking Your Learning—Possible Portfolio Entries***

Any of the activities included in Chapter 4 can be used as portfolio entries. Remember, the learning targets for this book are outlined in Figure 4.1, listed in Table 1.2, and described in detail in Chapter 1. The portfolio entry cover sheet provided on the CD in the file, “Portfolio Entry Cover Sheet,” will prompt you to think about how each item you choose for your portfolio reflects your learning with respect to one or more of these learning targets.

Specific recommendations for portfolio content follow.



## TRY THIS

### **Activity 4.6 Critique an Assessment for Good Design**

In Chapter 2 we introduced rubrics for analyzing and improving the quality of classroom assessments. The rubrics align with the learning targets in Table 1.2 and Figure 4.1. There is a rubric for each of the keys to quality classroom assessments—Clear Purpose, Clear Targets, Sound Assessment Design, Good Communication, and Student Involvement. In Chapter 2 we suggested it would be worthwhile to look at some sample assessments for the traits of Clear Purpose and Student Involvement. In Chapter 3 we proposed that it would be useful to examine sample assessments for the trait of Clear Targets.

Chapter 4 emphasizes the Sound Assessment Design key to quality classroom assessment. Therefore, once again, we recommend that it would be useful to examine at least one classroom assessment for the quality of the design—target-method match, sampling, and attention to possible sources of bias and distortion.

You will find the classroom assessment quality rubrics on the CD file, “Assessment Quality Rubrics.” In this case, we’ll be using the one called “Sound Assessment Design.”

Go through the same steps as previously:

1. Individually, read the strong level of the rubric first—“fast tracked.” Then read the beginning level—“side tracked”—and finally the middle level—“on track.” Mark the words and phrases that jump out at you as those really describing each level of quality.
2. Discuss any questions you have about the rubric with your colleagues.
3. Individually, look at your first sample assessment. If you think it is strong for the trait of Sound Assessment Design, begin reading the rubric at the “fast tracked” level. If you think it is weak, begin reading at the “side tracked” level. If you think it is strong, but “fast tracked” doesn’t quite match what you see in the assessment, read “on track.” Conversely, if you think it is weak, but what you read in “side tracked” doesn’t quite describe it, read “on track.”
4. There is no such thing as a “right” score, only a justifiable score. Justify your score using the words and phrases from the rubric that you felt described the assessment.

### Activity 4.6 (Continued)

5. Compare your scores to those of your colleagues. Discuss discrepancies and attempt to come to a resolution.

We have several sample assessments on the CD to use as defaults. See the file, “Assessments to Evaluate.” Our analyses of each of these samples are also on the CD in the file, “Assessment Critiques.”

When you finish analyzing your assessments, discuss your comfort level using the rubrics. Was it easier than last time? How easy was it to match the sample assessments to the rubrics? What questions does this process raise for you?



### TRY THIS

### Activity 4.7 Critique Your Own Assessment

Follow the procedure given in Activity 4.6 to apply the trait of “Sound Assessment Design” in the Classroom Assessment Quality Rubrics to an assessment that you currently use. Revise the assessment as needed based on the results.

Arter, Judy. and Chappuis, Jan. and Chappuis, Steve. and Stiggins, Richard. 2004. Assess How? Designing Assessments to Do What You Want. In CLASSROOM ASSESSMENT FOR STUDENT LEARNING: DOING IT RIGHT – USING IT WELL, 89-121. Portland, Oregon: Assessment Training Institute.

This reading is provided by the Assessment Training Institute, which provides comprehensive professional development, training materials and resources for teachers and school leaders so they can better manage the challenges of classroom assessment. This PBS Teacherline course on assessment and evaluation is only a beginning exploration of classroom assessment, therefore Teacherline encourages students to further develop their knowledge and skills in these areas by visiting the Assessment Training Institute’s Web site at: [www.assessmentinst.com](http://www.assessmentinst.com).

