\*\*Note: Remember to use Musial as reference.

Musial, D., Nieminen, G., Thomas, J., & Burke, K. (2008). Foundations of meaningful educational assessment. Boston: McGraw Hill.

CHAPTER 5 Selected-Response Assessments

Chapter Objectives

After reading and thinking about this chapter, you will be able to:

 • Understand why reliability and validity are essential characteristics of an effective test.

 • Describe the relationship between testing and assessment.

 • Explain where selected-response tests fit into assessment.

 • Describe the essential characteristics and assess varied examples of

 • Multiple-choice items

 • True-false or alternative response items

 • Matching items

 • Construct items that assess student understanding at multiple levels, from factual recall to understanding of principles and concepts.

Throughout this text, we have anchored our discussion of assessment with the image of a teacher sitting beside a learner, trying to understand what that learner knows or can do. The teacher may be asking the learner a set of questions on a specific topic or may be asking the learner to perform a task that will demonstrate mastery of a particular skill. This technique of using a set of focused questions or focused observations leads us to the concept of testing. Tests are part of the assessment process, and they can provide valuable information about the learner.

In this chapter we describe forms of tests and test items that require the learner to select an answer that has been predetermined by the teacher. We present procedures for developing, interpreting, and refining multiple-choice, matching, and true-false test items.

What are the foundational questions and issues underlying such forms of assessment? Think for a moment about the last test in which you invested quite a bit of study time. Now recall the test itself. How much of what you knew about the subject was actually covered on the test? When you create 107108a test for your students, you have identified those ideas and skills that you deem most important, but can any test fully capture the breadth or uniqueness of your students' understanding?

Foundational Questions for Your Consideration

 • Is an objective test (such as multiple choice, true-false, or matching) truly “objective”?

 • Since many of our schools' standardized tests rely heavily on true-false and multiple-choice items, is it proper to develop such instruments for classroom use for the purpose of preparing and familiarizing students with the format?

Testing and Assessment

As discussed in Chapter 2, testing is the process of evaluating students' knowledge or skills by their performance on a particular instrument or task that is presented in a controlled manner. Tests may take a number of forms, but in general they are intended to serve as an objective measure of learning. The type of questions that are used can vary widely in their formats (for example, true-false, essay, multiple choice). The environment can also vary (paper and pencil, oral, computer; timed or untimed; open notes or no notes). In addition, the number of questions in a single test can range from a small number of items focused on a single concept to multiple groups of items with each group leading more deeply into a concept. The type, precision, and depth of information obtained through a test depend on the question format, the number of questions per concept or skill, and the type of environment in which the test takes place.

Test results may also be used in a variety of ways. For example, a student's score on a test could tell us how well that student understands the causes of the American Civil War. We could also use that same score to measure how much the student has learned since he or she took an earlier test on the same subject. Or we could compare a student's knowledge to that of the rest of the class. And, taken together, the scores of all the students in the class on that same test can tell us something about the class's standing in their knowledge of the Civil War when compared to students across the state or nation.

108 109

Teachers and students are probably most familiar with the classroom-based selected-response variety of test, but in any school setting, many different types of tests are used. Physical education classes, for example, may require that students demonstrate proficiency in particular skills, such as the time it takes them to run a mile. In a history class, students may be asked to recite the preamble to the Constitution.

Testing is one form of assessment, which we defined in Chapter 1 as the art of placing learners in a context that clarifies what it is a learner knows and can do, as well as what a learner may not know or be able to do. But assessment encompasses many more ways of evaluating students, both formally and informally, than testing alone. In addition to testing, assessment includes methods for evaluating student products, work habits, complex behavioral skills, abilities, dispositions, attitudes, and interests. Further, assessment focuses not only on student performance and potential but also on the factors that influence a student's learning—personality, motivation, home environment, and areas of exceptionality, for example. Assessment tells us what our students have learned and their capacity to learn. Assessment gives us evidence for reflection about our own learning and teaching. Assessment is the basis for all critical decisions about our students.

The purpose of this book, then, is to help you learn to create tests and other formal assessments that show how much students have learned and also to help you capture and communicate less formal indicators of student learning and development.

In summary, testing can be used to:

 • Show the depth of understanding of an idea or mastery of a skill.

 • Show a student's growth over time in a particular area of knowledge or skill.

 • Compare one student's or one group's achievement to another's on the same task.

 • Predict students' future performance.

Ask Yourself

Think about some of the tests you have taken in the past—math tests, driver's license tests, hearing and vision tests. Each of those tests measures or assesses a particular skill—your understanding of algebra, your ability to safely operate a car, the acuity of your eyesight. But in what ways might such tests be limited? Are they truly the most accurate means to assess such skills? Can you think of a more comprehensive way to assess those skills or characteristics?

109 110

Validity and Reliabity in Testing

Before considering how to construct sound test items, try to answer the following question: What is a good test? You might answer this in many different ways, but from teacher-made midterms to international tests of mathematics, we strive to develop tests that are both consistent in the way they measure learning and accurate in what they measure.

Good tests—tests that are reliable and valid—give us results that we can use to make decisions about students. Recall from Chapter 1 that validity is what we have in mind when we ask the question, “Am I testing what I think I'm testing?” And reliability refers to the stability or consistency of the test results. Let's consider reliability first.

Test Reliability

When you give a test to your students, you would like to think that their responses are stable. That is, you would hope that their scores would be basically the same if you had given the same test at another time or place (assuming the students had the same amount of instruction and practice). As a teacher, you want to avoid introducing error into the testing situation. You want the students' scores to reflect what they know—to be as close as possible to their true level of knowledge and skill, with as little error as possible.

At the same time, it is important to realize that there is always error in any kind of testing or measurement. In fact, it can be helpful to think of a student's test score as a combination of two factors: the true score and some error. Some of the error comes from the student. If the student is tired, ill, or bored, these factors can have the effect of lowering the true score. We say that the student “didn't do her best.” Or if the student is guessing at the answers, his test score will also contain error that raises or lowers his score. On the other hand, some error is external to the student, and most of this is under the teacher's control. Factors such as poorly worded test items, confusing directions, and classroom interruptions during the test can also lower a student's true score. Of course, the true score is an abstraction, a hypothesis about what the student is truly capable of.

Reliability is important both to teacher-made classroom tests and to standardized, norm-referenced tests such as the SAT. We will discuss standardized tests in Chapter 12 and will describe how to statistically determine a test's reliability. In the classroom, however, you will most often be using your professional judgment rather than statistical calculations to determine if your students' scores are reliable.

Types of Reliability

There are several ways to determine a test's reliability, each of which has a different purpose and focus. Test-retest reliability is a means of determining 110111reliability that is easy to understand and apply in a classroom setting. Test-retest determines a test's reliability by administering the same test to the same group of students at two different times and then comparing the scores on the first and second testing. Comparable scores between the two testings indicate stability or consistency over time.

There are several cautions with test-retest reliability, however. First, if the students have had the opportunity to learn some additional information that is relevant to the test contents, then you would not expect stability in the test scores. Second, the time between administrations can also influence the test's reliability. If the testings are too close together, then familiarity with specific items can influence the score on the second administration.

Another type of reliability is important when a tester uses two different forms of a test and wants to be certain that they are equivalent. Equivalent forms reliability determines if two forms or versions of the same test contain items that are comparable in form, length, and difficulty. Equivalent forms reliability is more common in standardized testing, where tests often have two or more forms. In the classroom, you might create a second test similar to one that you have already given in order to allow a group of students who missed the first test to make it up.

Internal consistency reliability is relevant when you are creating a test that is focused entirely on one concept or specific area. In this case, you are interested in making sure that all of your test items are aimed only at that concept or specific area and that there are no unrelated items within the test.

Finally, inter-rater reliability refers to the degree of consistency of scoring when student responses (usually an essay or constructed-response item) are assessed by more than one scorer. Tests that use selected-response items (multiple choice, true-false, for example) require little judgment in scoring—all you need is an accurate scoring key. But essay items may elicit opinions or guesses from the person doing the scoring. We will discuss how to minimize error due to lack of agreement between raters in Chapter 6.

Validity

As we suggested, reliability can be determined with reasonable confidence using various strategies and statistics. But is a test that is reliable also valid? Not necessarily. In fact, a test can be very reliable and not at all valid. What might happen if, in a psychology class, the teacher administered a 50-item, multiple-choice biology test? Some students would do quite well, others would do not as well, and some would do a lot of guessing. If the teacher administered the same test (test-retest reliability), individual students' scores probably would not differ by much. So this 111112test is reliable, but it in no way measures what the teacher intended to measure.

This is an absurd example, but it points to ways to think about validity in classroom tests. Content validity refers to the degree to which your test aligns with what you intend to measure, which would usually be your teaching objectives. For example, if you administer a test intended to cover the first 8 weeks of mathematics instruction, but the items only address concepts covered in the first 4 weeks, it does not have strong content validity.

Criterion validity asks how a particular assessment aligns or correlates with another measure. Does your classroom-level third-grade reading test, for example, correlate with districtwide assessments of third-grade reading?

Validity should be considered as you construct or select your classroom assessments. Careful planning of the test, consideration of individual test items, and a determination of exactly what it will cover are essential to developing a valid test. The planning involves comparing the list of learning objectives you intend for students to master to the list of test items you have constructed.

Validity becomes especially important when considering state-level learning standards. And state standards should provide a meaningful framework in the construction of such items. Ask yourself, as you construct an item or test, “Does this item indicate understanding relative to the state standard in this content area?”

Validity is not a precise practice but consider the implications of administering and then interpreting an invalid test. In the previous example, we thought we created a test that covered 8 weeks of math concepts but somehow we only included items from the first 4 weeks. When we interpret the scores on this test, we are going to make serious mistakes about our understanding of students' progress. Figure 5.1 summarizes ways to increase a test's reliability and validity.

Figure 5.1 How Can You Increase Reliability and Validity in Classroom Tests?

112 113

Ask Yourself

Recall a time when you took an unscheduled test and received a poor score. Did you feel that the score was a true measure of your knowledge? Why or why not? Did the test results help you study further so that you could do better the next time? What could have been done differently to make the experience more effective?

Where Do Selected-Response Tests Fit into Assessment?

A selected-response test is one tool that allows teachers to tap certain kinds of student achievement with accuracy and efficiency. Because of the level of detailed planning they require in order to be valid, selected-response tests also challenge teachers to clearly specify the important content and skills that they have been teaching and that they want to include in a test.

Selected-Response Tests

Chapters 5 and 6 each focus on one category of paper-and-pencil assessment. In this chapter, we will be presenting selected-response items. These are items in which students are not asked to produce an original answer because the answer to each question appears in some place on the test. In multiple-choice, matching, and true-false tests, the teacher has predetermined the correct answer; and, while students may need to perform a simple calculation for a math problem, for example, they are not expected to produce an answer that is not already shown on the test itself. In Chapter 6, we will present another type of paper-and-pencil assessment in which the responses are student-produced. These assessments include simple items like fill-ins and short answers as well as more complex essay items, all of which require students to generate an answer that is not supplied by the test.

One note: We use the term “paper-and-pencil test” in this text knowing that, in many academic settings, such assessments are often carried out using classroom computers. We use the term more generally, however, to describe an assessment that includes teacher-made items as described in this and the next chapter, regardless of how the test is administered.

Advantages of Selected-Response Tests

There are plenty of reasons that teachers would want to use selected-response tests. Good selected-response tests

 • Provide formative feedback about particular strengths or areas needing improvement.

 113 114

 • Provide diagnostic information about student errors and misunderstandings.

 • Yield scores that are easily summarized for parents or administrators.

 • Allow for reasonable comparison among students and between classes.

 • Are relatively easy to grade.

 • Prepare students for some of the context and format of the standardized testing that they will encounter throughout their schooling.

Meaningful selected-response items can assess low-level knowledge through higher-order cognitive skills. Different item types offer different strengths in assessing learning. The construction of such items demands attention and deliberate thought.

Disadvantages of Selected-Response Tests

Selected-response items alone do not offer insight into students' reasoning. Good multiple-choice items have one right answer, and that is how the students' scores are determined by how many right answers they choose. But it may be just as revealing for you as a teacher to know why your students chose a particular answer as it is to know their test score. One way to achieve this is to allow students an opportunity to explain their incorrect answers to you after they have seen their scores.

The most frequent criticism of selected-response test items is the possibility of students guessing the correct answer. This is a valid criticism, although it only applies to lucky guesses. Students who correctly guess some answers will receive a test score that does not accurately reflect their level of understanding.

Ask Yourself

Recall an instance when you were proud of your performance on a test. What type of test was it? Objective? Constructed response/essay? Why were you particularly proud of your performance on that test?

True-False or Alternative-Response Items

For many students and teachers alike, true-false items (also called alternative-response or binary-choice items) seem not to be the most effective way to assess or demonstrate understanding—and certainly not the kind of deep understanding that we hope to develop in students. Further, isn't it true 114115that students can simply guess on true-false items? Both of these criticisms may be valid, but as you will see, true-false items can, in fact, assess learning in a meaningful way.

Digging Deeper What Did the First Standardized Test Items Look Like?

How far have we come in the development of ways of assessing student learning in an objective manner? Alfred Binet was commissioned by the French Ministry of Education to develop a test that would identify students who needed remedial attention in school. The result was one of the first tests of intelligence, and Binet's influence is still recognized in the form of a widely used intelligence test that bears his name: the Stanford-Binet.

The following task is taken from Binet's “New Methods for the Diagnosis of the Intellectual Level of Subnormals” (1905), which was intended to identify degrees of impairment and levels of ability. As you read through this task, ask yourself these questions: How objective were these original tests? Do the instructions to the test administrator allow for too much interpretation of the child's answer? What might be the consequence of such testing practices? Do you think a more objective test item type could give us a more realistic picture of a student's true ability?

Comparison of Known Objects from Memory

This is an exercise in ideation, in the notion of differences, and somewhat in powers of observation. Procedure. One asks what difference there is between paper and cardboard, between a fly and a butterfly, between a piece of wood and a piece of glass. First be sure that the subject knows these objects. Ask him, “Have you seen paper?” “Do you know what cardboard is?” Thus ask him about all the objects before drawing his attention to the difference between them. It may happen that little Parisians, even though normal, and eight or nine years old, have never seen a butterfly. These are examples of astounding ignorance, but we have found, what is still more extraordinary, Parisians of ten years who have never seen the Seine.

After being assured that the two objects to be compared are known, demand their difference. If the word is not understood, take notice and afterward choose more familiar language. “In what are they not alike? How are they not alike?” Three classes of replies may be expected. First, that of the children who have no comprehension of what is desired of them. When asked the difference between cardboard and paper, they reply, “The cardboard.” When one has provoked replies of this kind, the explanation must be renewed with patience to see if there is not some means of making oneself understood. Second, the absurd replies, such as, “The fly is larger than the butterfly,” “The wood is thicker than the glass,” or “The butterfly flies and so does the fly.” Third, the correct reply.

Sources: Binet, 1905; Plucker, 2007.

Compared to other types of selected-response items, true-false items are fairly easy to write and are also easy for teachers to score—the answer can be only correct or incorrect. It is this apparent simplicity of construction and scoring that has probably led us to believe that true-false items cannot be powerful, discriminating questions. But they can be if they are properly written. Figure 5.2 presents some of the advantages of true-false tests.

115 116

Figure 5.2 What Are the Advantages of True-False Items?

What Does a Good True-False Item Look Like?

A true-false item begins with a proposition—a statement that asserts a particular truth about an idea, relationship, or concept. Think about the following:

 1. The earth experiences four different seasons. (T)

 2. The earth travels in a highly elliptical orbit around the sun. (F)

 3. The earth's seasons are caused by changes in the distance of the earth from the sun. (F)

 4. The earth's seasons are caused by changes in the earth's tilt, which changes the directness of the sun's rays. (T)

The items are stated as propositions—statements of fact that require recall of a fact (“The earth has four seasons”) or a relationship (such as between the changing distance from the earth to the sun and its effect on the seasons).

These examples demonstrate several important principles about true-false items:

 • First, they can test student learning from simple recall to higher-order understanding of principles.

 • Second, these items are short statements from which a variety of items can be developed from a single idea or concept.

 • Third, statements two and three are presented as false statements that are actually typical of misconceptions often held by students (and teachers, too), another way in which true-false items can assess meaningful learning and lead to better teaching.

These items respond to a common criticism of true-false items: that they typically only assess students' recall of facts and encourage rote memorization as a learning strategy. As a teacher, there will be times that you will want to assess simple recall, such as on a quiz that is intended to determine 116117whether students have completed their assigned reading. But if you have identified clear objectives and learning targets for your students, true-false items can also be an effective way for you to assess understanding of important propositions.

Guidelines for Effective True-False Items

What makes an effective true-false item? As you develop your true-false items, keep the following guidelines in mind.

 1. True-false items should focus on an important idea rather than a trivial fact. Look at the following two items and ask yourself which contains the more important idea:

 • Jean Piaget's theory of development has four stages. (T)

 • A child in Piaget's concrete operational stage is able to think inductively.(F)

 The first of these two questions is really only important in that students should probably be expected to be able to name and describe the stages, and recalling that there are four may help in remembering the stages themselves. But the second question gets at several important ideas that are essential to understanding Piaget's developmental theory. Whether a student can simply recall that Piaget's theory has four stages is less likely to be worth a test item than whether the student has an understanding of the cognitive characteristics of each stage. Students can also, through their answers to well-written true-false items, demonstrate understanding of important ideas through application of that idea. Consider the next two true-false questions.

 • Vygotsky believes that cognitive development occurs in age-related stages. (F)

 • A biology teacher who administers a pretest on the first day of class to determine students' knowledge of biology is assessing his students' zone of proximal development. (F)

The first question may be somewhat important in that it makes a critical distinction between Vygotsky and other theorists, but the second question is more meaningful for two reasons. First, it assesses students' understanding of a central concept in development, and, second, it expects students to apply their understanding to a real-world context, in this case, a classroom.

2. Answers to true-false items must be either true or false. This seems self-evident, but think back on tests you have taken. How often have you or a classmate argued with a teacher about a true-false item that might have been correctly answered as either true or false? 117118Another charge against true-false items is that they are inherently ambiguous, when, in fact, the right answer should always be defensible. Read the following items:

 • Students' self-efficacy in mathematics is positively correlated with persistence in mathematical problem solving. (T)

 • Self-esteem is related to higher academic achievement. (F)

The first of these questions points to a clear relationship between two different variables—self-efficacy and problem solving. It is clearly worded, and, although it is not stated in absolute terms, the question is much more true than false. The second question could be argued either way. On the one hand, higher self-esteem may be slightly positively correlated with achievement, suggesting that the item should be marked true. At the same time, there are many factors that are more strongly related to achievement, suggesting that the item should be marked false. An item like this opens the door to arguments between students and the teacher.

3. A good true-false item avoids the quotation of phrases from the textbook. It is sometime tempting for teachers to create test items by repeating statements from the textbook after changing a word or two. This practice has perhaps led to some of the criticism about true-false items. For example, compare these two items:

 • A person's score on an intelligence test was originally calculated as a ratio of mental age to chronological age and termed the Intelligence Quotient. (T)

 • Children with a high IQ typically perform better on standardized tests than children with a lower IQ. (T)

The first question may initially seem useful in assessing students' understanding of intelligence, but it has several problems. The question is written using language that sounds very much like textbook phrasing. In other words, students may recognize that the question is not asking for understanding but rather recall from their reading. The second question, however, expects students to understand the relationship between IQ and other important skills.

4. A good true-false item does not contain inadvertent clues. Teachers sometimes will insert a word or two into a true-false item in order to throw off the test taker. Very often, this is done by making a positive statement into a negative statement as in these two items:

 • B. F. Skinner argued that learning is characterized by a response followed by a reinforcing stimulus. (T)

 • B. F. Skinner argued that learning is not characterized by a response followed by a reinforcing stimulus. (F)

118 119 The false statement is not only false, but it will immediately appear to students as false because of the unnatural placement of the word not in the middle of the statement. Such a practice is a form of laziness or lack of attention to creating a good test item. It is quickly recognized by students, who are then given the opportunity to be lazy themselves and use otherwise irrelevant cues in the item rather than depending on their understanding to figure out the correct answer.

Ask Yourself

Get a copy of a test that has a number of true-false items. Examine the characteristics of a good true-false question and determine which of the items has all of these characteristics. What do you conclude about the relative merit of these true-false items in assessing your understanding about the topic?

Multiple-Choice Items

Multiple-choice items are perhaps the most widely used type of test item in schools, for several reasons. Although they are not as easily constructed as true-false items, they are easy to score, and they can be written to assess various levels of understanding. Also, multiple-choice tests are less susceptible than true-false tests to the effects of guessing. Multiple-choice items are the type students are most likely to encounter on standardized tests, so they have come to expect the familiar multiple-choice format. Figure 5.3 summarizes the advantages of multiple-choice items.

Figure 5.3 What Are the Advantages of Multiple-Choice Items?

The construction of an effective multiple-choice item appears simple but requires care and precision on the part of the teacher who uses it. A multiple-choice item has three components:

 • The stem, which presents a question or an incomplete statement to the student

 • A single correct or best answer

 • Distracters, usually three or four per question, which present plausible, alternative answers to the correct answer

You will find as you begin writing multiple-choice items that it is relatively easy to develop a meaningful question, a right answer, and one or two plausible distracters. The difficulty often comes in trying to develop enough meaningful distracters to assure that the item truly tests what you intend.

The Stem

Let's begin with the stem. The two most common ways to present a multiple-choice question to a student are a direct question that the student must answer or an incomplete sentence, as illustrated in these two examples:

 • What is the capital of Alaska?

 • Alaska's capital is named \_\_\_\_\_\_.

Either of these common stem formats is acceptable, although some texts recommend that you write the stem in the form of a direct question because it can present the student with a clearer task. Regardless of the form of stem, it is important to keep the wording as succinct and clear as possible. All of the words in the stem should be relevant to the task.

At the same time that you are writing a stem that is short and to the point, you also must be sure that your stem presents the question fully and completely. One useful way to think about your stem construction is to ask yourself if you could answer the question even without looking at the answer and distracters. Consider this stem:

 • Chicago is \_\_\_\_\_\_\_.

Without an answer or distracters, there are quite a few ways you could answer it: Chicago is a city, it is the largest city in Illinois, it was the site of a great fire in 1871, it was the subject of a Carl Sandburg poem, and there are many other possible answers. Each of these answers is correct, but the test writer probably had only one of these answers in mind when writing the test. A better way to pose the stem would be:

 • Which Illinois city has the largest population? or

 • The Illinois city destroyed by a great fire in 1871 was named \_\_\_\_\_\_\_\_\_.

As you construct your multiple-choice items, consider carefully the relationship between the stem and the answers. Grammar and punctuation are important in creating a good test item, one that is understandable and at 120121the same time gives no accidental hints to the student. Take a look at the following items.

 • Which of the following is a mammal?

 a. cat

 b. fish

 c. turtle

 d. bird

Notice that in this example, the item is expressed as a direct question. The stem begins with a capital letter and ends in a question mark. The answers do not need to complete a sentence and are not proper nouns. Therefore, the correct answer and the distracters begin with lowercase letters and have no punctuation following.

Now consider another type of item.

 • A portion of land nearly surrounded by water and connected with a larger body is called

 a. an island.

 b. a peninsula.

 c. an isthmus.

 d. a land bridge.

This item is stated as an incomplete sentence, so in this instance the stem ends without punctuation. (You may also complete the stem with a colon.) Notice, though, that the answer and distracters, which serve to complete the sentence, begin with lowercase letters and have terminal punctuation (a period).

The stem should also ask a question for which there is either one correct or best answer. What is the difference in the following two items?

 • Which U.S. city has the highest literacy rate?

 a. New York

 b. Seattle

 c. Los Angeles

 d. Baton Rouge

 • Which of the following U.S. cities is most important to international trade?

 a. New York

 b. Seattle

 c. Los Angeles

 d. Baton Rouge

The first question can have only one correct answer. Even if the student does not know whether the answer is New York, Seattle, Los Angeles, or 121122Baton Rouge, there is only one, single best answer. The second question, however, does not have an obvious best answer. You would probably find that experts on the topic would differ in their answers to the second question. This kind of item is asking the student to respond with an opinion. While interesting, such questions are not suited to a multiple-choice format. They are better suited to a constructed-response or essay test. Here is another example:

 • The most significant maritime event of the twentieth century was

 a. the sinking of the Titanic.

 b. the sinking of the Eastland.

 c. the grounding and oil spill of the Exxon Valdez.

 d. the attack on Pearl Harbor.

Again, this question is more appropriate to a discussion or an essay question.

Presenting a direct or indirect question is not the only way to create the stem for a multiple-choice test item. An effective way to vary the complexity of the task represented by the multiple-choice item is to vary the format of the stem. For example, for younger students, the teacher may read the stem aloud, while the students look at the answers and mark one of them. Or a series of questions could be focused on a passage, read by the student, that states a problem. Or the stem could include graphics that require the student to interpret them in order to answer the question. You will find it helpful to consider the possible array of stems when you begin to design a multiple-choice test.

Distracters

Distracters are the feature of multiple-choice items that allow teachers to make generalizations about student learning. Therefore, the distracters must be constructed so that they require understanding and careful thought but are sufficiently wrong that students who are prepared for the test can identify which is the correct answer. On the other hand, well-written distracters should all appear plausible to students who are not prepared. As you write your multiple-choice test, there are a few important considerations in developing your distracters.

Plausibility First, distracters should be plausible but incorrect. Consider the following:

 • Which of the following is the best predictor of college grade point average?

 a. a student's performance on the SAT

 b. a student's performance on the Presidential Physical Fitness Test

 c. a student's height

 d. a student's weight

122 123

Resource for Your Assessment Toolkit Formatting Options for Multiple-Choice Items

Multiple-choice items have a variety of formats. The stem, the distractors, and the correct answer can have different forms, and these differences can enhance the effectiveness of an item to assess student learning. It is important to examine the structure of your items in light of the type of instruction you have given your students. For example, it would be inappropriate to test students using a multiple-choice stem that contained a long example of a concept when all you have taught students is the definition of the concept.

This list provides specific names for the different formats of multiple-choice items. You may find it useful to review these examples when you are developing a multiple-choice assessment for your students. (The correct answer is starred.)

Stem Variations for Multiple-Choice Items

Oral Stimulus Teacher says, “Which is the number four hundred and three?” The student sees only choices, not stems.

 a. 34

 b. 43

 c. 304

 \*d. 403

Passage Related Kim has a 5-dollar bill and wants to buy as many 41-cent stamps as she can. How many will that be and what change will she get?

 1. To solve this problem, what would be the best operation to use?

 a. addition

 b. subtraction

 c. multiplication

 \*d. division

 2. How can you find out how much change Kim will get?

 a. Subtract the number of stamps from $5.00.

 b. Add the cost of the stamps to $5.00.

 \*c. Find the remainder after dividing $5.00 by 41cents.

 d. Multiply the number of stamps by 41 and add that to $5.00.

Answer Variations for Multiple-Choice Items

Five students have the following scores:

Kelsey

15

Jason

10

Jennifer

18

Greg

13

Yasmin

14

123 124

For each of the situations below, tell whether the change would cause the mean to …

List Reference

 a. increase

 b. decrease

 c. stay the same

 d. can't tell

Although there may be research out there that somehow relates height or weight to college grade point average, answers b, c, and d are really so implausible that every student should answer this question correctly. Three implausible distracters make for a poor test item, but even one distracter that is not plausible reduces an item's ability to discriminate between students who know the material and those who do not.

All of the Above, None of the AboveIt is also tempting to include one or two distracters that imply that none, two, or more answers are possible. Items that offer “all of the above,” “none of the above,” or other similar choices as answers have several potential drawbacks. First, some students will recognize the first answer as correct and fail to read the other possible answers, never getting to the “all of the above” choice. Second, some teachers will only use “all of the above” or “none of the above” when they are unable to come up with a reasonable third or fourth distracter, and students will come to recognize this pattern and discount those choices as possible correct answers. Or, conversely, ­students might recognize that teachers only use such distracters when they are in fact the right answer. If you decide to use these, do so both 124125as correct and as incorrect answers. Further, when using “all of the above” as an answer, it is important to ensure that all answers are entirely correct.

Indirect Clues Another common error in writing multiple-choice items is that teachers sometimes provide indirect clues to an answer in either the stem or among the answers. What clues does the stem give you about the answer to this multiple-choice question?

 • Sternberg's three-part model of intelligence is known as the

 a. triarchic model.

 b. multiple intelligence theory.

 c. information processing model.

 d. IQ.

Even students not familiar with Sternberg or his theory of intelligence could guess that “a” is the correct answer to the question, due to a linguistic clue. Since his theory has three parts and answer “a” has the prefix tri in the first word, students might reasonably guess the answer to this question.

Students are also aware that answers that are longer or constructed differently from the other answers often represent the correct response. As you write your test, it is important to ensure that the correct answer does not appear to be written with greater precision or clarity than the distracters. Often such attention to detail results in the correct answer being much longer than the distracters. Here is an example:

 • What is the most widely accepted definition of intelligence?

 a. IQ

 b. a source of human difference

 c. the level of mental development

 d. the combination of verbal ability, problem-solving skills, and the ability to adapt to and learn from life's everyday experiences

To write good multiple-choice items, you will need practice. You will also need to review your items for possible faults that might give awaythe answer or make the item impossible for students to answer correctly.

The Correct Answer

It probably seems obvious that a right answer is a correct answer. However, it is useful to reflect on the precise characteristics of a correct answer. One characteristic of the correct answer relates to the concept of “best.” An answer is correct when it is the best choice from a list. “Best” implies that there are other choices that have a reasonable connection to the question, but one 125126choice has the most important or appropriate connection. For example, you may wish to assess if a student understands the concept of a conflict. The question might be worded as follows:

Resource for Your Assessment Toolkit Common Errors in Constructing Multiple-Choice Items

Here are some examples of multiple-choice question errors that are often made by practicing teachers. Each error has a specific phrase that will help you remember to avoid the error when you're developing multiple-choice questions. Review this list when you write a multiple-choice item and check your item against these common errors. (The correct answer is starred.)

Item with Vague Stem

 • George Washington

 \*a. was a great general.

 b. was born in 1776.

 c. wrote the Declaration of Independence.

 d. abolished slavery.

Double-Barreled Item

 • In the past, whaling and cotton growing were important industries in

 a. the Northeast and the West.

 b. the Middle West and the Southeast.

 \*c. the Northeast and the Southeast.

 d. the West and the Middle West.

Item with Multiple Correct Answers

 • Two of the most important characteristics of a good test are validity and

 \*a. reliability.

 b. accuracy.

 \*c. consistency.

 d. logic.

Grammatically Non-Parallel Item

 • The correct way to find the area of a triangle is to

 \*a. multiply the length of the base by the height and take half.

 b. add the lengths of the sides.

 c. area = base × height.

 d. 1/2 (side 1 + side 2 + side 3).

Inconsistent Alternatives

 • The difference between an achievement test and an aptitude test is

 a. the test is objectively scored.

 \*b. that one measures what a person has already learned and one measures the person's ability to learn new material or skills.

 c. between power and speed.

 d. between personality and cognition.

Implausible Distracters

 • What is a denominator?

 a. the top number of a fraction

 \*b. the bottom number of a fraction

 c. a kind of church

 d. all of the above

Trick Alternatives

 • What is 3/4 + 3/8?

 \*a. 9/8

 b. 6/12

 c. 6/8

 \*d. 1 1/8

 • Which of the following characteristics best defines the concept of conflict?

 a. a disagreement between two people

 \*b. a struggle between opposing interests

 c. competitive sports like football

 d. mediation between parties

In this example all of the answers relate in some way to the concept of conflict, but “b” is the best answer. Choice “b” states the underlying characteristics that could relate to many types of conflict, while “a” and “c” illustrate the concept rather than defining it. Choice “d” is only slightly related.

Another way to think of the correct answer is that it is the choice from among other choices that most closely matches the objectives and content of the instructional unit. If you, as teacher, choose to emphasize a specific aspect of some content, then the correct answer will include that emphasis. For example, if you developed an instructional unit that emphasizes that all life cycles have distinct, developmental phases in the life span of an organism, then the correct answer for a question about the key characteristic of a life cycle would be the response that includes distinct developmental phases.

Finally, when considering the characteristics of the correct answer, caution is in order. It is important to examine the distracters for a test item and eliminate possible overlap with the correct answer. A distracter should relate to the question in some way to truly test the student's knowledge. But you must make sure that the distracter's relationship to the question is distinct from the correct answer's relationship to the question. Again, be aware that the wording, the emphasis, and the relationship between the correct answer and the distracters is a complex task that requires a good deal of reflection.

The Added Value of Multiple-Choice Items

As we have suggested, multiple-choice items are of practical value to teachers and students for several reasons. First, they are relatively easy to construct and very easy to score. They also represent the most common item type on standardized tests such as the ACT and SAT. Now let's look at a third advantage: their ability to assess a variety of learning outcomes and levels of understanding.

For example, you may want to assess students' recognition of characters in a novel, ability to define vocabulary terms, or recall of state capitals. Or you may want to dig deeper and assess whether students understand 127128the effects of the jet stream on weather patterns. Multiple-choice items can be constructed to assess a range of learning, from recall to understanding of processes and principles. The following sections provide some examples of such purposes for multiple-choice items.

Recognition of Terms and Vocabulary One of the simplest learning outcomes is the recognition of specific terms and vocabulary that provide the basis for further understanding. For example, your sixth-grade class is studying a new science unit, and, to ensure that they are able to recognize and define key terms, you develop the following items.

 • The study of the interrelationships of living organisms and their environment is

 a. ecology.

 b. astronomy.

 c. physiology.

 d. biology.

 • An organism living in or on another organism is

 a. a predator.

 b. prey.

 c. a parasite.

 d. a host.

Factual Knowledge Factual knowledge is essential to the development of deeper conceptual understanding. People who are experts in their fields demonstrate an ability to look at a problem from many perspectives, but their understanding is based on extensive factual knowledge. Multiple-choice items can be written to assess students' grasp of ­discipline-based factual knowledge, such as the following geography questions.

 • Which of the following states does not border Oklahoma?

 a. Colorado

 b. Missouri

 c. Nebraska

 d. New Mexico

 • Which is the smallest of the Great Lakes?

 a. Lake Erie

 b. Lake Huron

 c. Lake Michigan

 d. Lake Ontario

 e. Lake Superior

128 129

Procedural Knowledge Factual knowledge provides evidence that students know who, what, where, and when. Procedural knowledge, however, provides evidence that students know how—how something works or how one should approach a problem or situation. Before we allow students to use an expensive or potentially dangerous piece of laboratory equipment, for example, we might develop multiple-choice items that assess procedural knowledge.

 • The correct procedure for combining acid and water is to

 a. add acid to large amounts of water.

 b. add water to large amounts of acid.

 c. add acid to water, cool, and swirl.

 d. add water to acid, cool, and swirl.

 • If you are the first person on the scene at a car accident, what should you do first?

 a. Call 911.

 b. Assess the scene to see if it is safe.

 c. Check victims for signs of pulse and breathing.

 d. Move the victims away from the site of the accident.

Higher-Order Thinking When presented with propositions, conditional information, or hypothetical situations, students can be assessed on the degree to which they can apply their learning or transfer understanding to new situations. Multiple-choice items dealing with cause-and-effect relationships, for example, can assess students' understanding of the relationship between facts.

Consider how the following examples move students beyond simple recall.

 • Your students take a quiz every Monday over an assigned reading. The results of your students' first several quizzes were very low, so you begin praising students who scored well on the quizzes, and over time your class's quiz performance began to improve. This is an example of

 a. operant conditioning.

 b. observational learning.

 c. information processing.

 d. social constructivism.

 • Which of the following is the best example of the principle of surface tension?

 a. a windshield wiper wiping away rainwater

 b. a baseball leaving the barrel of a bat

 129 130

 c. an insect walking on water

 d. a braking car's tires leaving tread marks

 • You fire a gun straight ahead of you and drop a ball at the same time and from the same height as the barrel of the gun. Which object will hit the ground first?

 a. the bullet from the gun

 b. the ball

 c. neither—they will land at the same time

Multiple-choice tests are versatile and valuable indicators of student learning, but effective multiple-choice items require planning and thought in their construction. As with any skill, you will refine your ability to write effective items over time. Just as important, you will begin to recognize characteristics of effective items on multiple-choice items constructed by others.

Ask Yourself

When you are a teacher, your students will probably ask you what kinds of items they can expect to see on an upcoming test. Now that you have had a chance to explore common item types in some depth, how would you answer your students? Do you think it is useful to let them know the purpose of the assessment (for example, will they be expected to remember names and dates rather than understand principles)? Do you think that knowing what types of items will be on a test leads students to study differently?

Matching

Matching items ask students to identify an item in one column with a closely associated item in a second column. In other words, in matching items students are not expected to create new information in order to answer correctly but rather to be able to identify common characteristics between two sets of items similar in nature.

Advantages and Disadvantages of Matching Items

Like multiple-choice and true-false items, matching items are relatively easy to construct and to score. One limitation of matching items is that they cannot reasonably assess learning beyond factual recall. Although they are not as discriminating as true-false or multiple-choice tests, they can assess a broad array of factual learning relatively quickly.

130 131

Creating Good Matching Items

There are steps you can take to ensure that the matching items you develop are meaningful for the kinds of factual recall you intend to assess.

Consider this question. If you have a list of four facts about U.S. presidents in one column, do you have four names of presidents in the second column? Not necessarily.

Which president was associated with each event?

 1. inaugurated as president in New York City

 2. established the U.S. Forest Service

 3. president during the ratification of the Fourteenth Amendment

 4. elected to four terms as U.S. president

Presidents

 a. John Adams

 b. Thomas Jefferson

 c. Abraham Lincoln

 d. Franklin D. Roosevelt

 e. Theodore Roosevelt

 f. George Washington

Offering an unequal number of items in your two columns is an effective way to reduce the possibility of answering correctly through the process of elimination. For this reason, it can be useful to qualify your instructions with the statement that items may be used more than once or not at all. In the example above, if only four presidents had been listed for four events, students who could make three correct matches would know the fourth match was the only combination left.

Another important consideration in the development of matching items is that the items in your columns be grouped homogeneously. For example, think about the homogeneity of the following matching exercise regarding famous artists.

Artist

 1. Leonardo DaVinci

 2. Edward Hopper

 3. Michelangelo

 4. Auguste Rodin

 5. Grant Wood

The column at left seems rather diverse with artists from very different times and places. Perhaps the teacher's purpose is to assess general knowledge of art as a pretest. Or she may be working with younger students who may only have familiarity with a limited number of major works of art covered in a textbook. But, for older students or students with a deeper understanding and familiarity with art history, the above matching exercise 131132would not be appropriate. For these students, it may be more appropriate, for example, to match a particular artist or work of art with a period in art history, such as impressionism or the Renaissance.

You might also incorporate images or pictures into a matching exercise, for example, by having students match the outline of a U.S. state with the state's name. Also, since we read in English from left to right, matching items should be written with the longer phrases in the left-hand column and the shorter responses (names, dates) in the right-hand column.

Finally, for ease of administering the test as well as scoring, you can make the task of matching much less cumbersome by keeping all of your pairs of items on one page and by keeping your pairs of items in a logical order. For example, if one of your columns includes a list of important dates in U.S. history, the dates should be arranged chronologically. If the column is a list of names, you may want to put them in alphabetical order.

Ask Yourself

Now that you have had a chance to think about the potential value of selected-response items, do you believe that the tests you have taken as a student used such items to assess deeper levels of learning? Do you believe that students and teachers recognize that true-false items are effective in assessing higher-order thinking? Next time you take a test or see sample items in a textbook, try to evaluate the items using the criteria presented in this chapter.

Recommendations for Developing Selected-Response Items

The development of good selected-response assessments is a skill that requires careful attention to item construction. Improvement in item construction comes with experience, practice, and reflection. Over time, you will become more adept at writing items and aligning items with student ability and learning targets, and you will also become more critical readers and consumers of items developed by others, such as those found in test banks and teachers' manuals. With that in mind, we conclude this chapter with two additional strategies that will aid you in the development of selected-response assessments, regardless of item type.

 • Share your tests with colleagues. Regardless of your skill in constructing selected-response assessments, the eyes of a colleague can give you an indispensable perspective. Have you taken tests during which several 132133students raise their hands to ask the teacher to clarify a question? Keep in mind that you want all of the students to read each question the same way, and the best way to make sure of this is to ask a colleague—preferably one teaching the same subject—to read the test items carefully for clarity.

 • Watch spelling and punctuation. Tests can engender anxiety among some students. For some, the appearance of a misspelled word may trigger thoughts that the teacher is intentionally misleading students. Also, students should expect to encounter items without distraction, and spelling and punctuation can sidetrack students' focus. Proofread your items carefully.

Ask Yourself

Sharing your work with others is a powerful way to self-assess. It can be difficult to receive constructive criticism from colleagues or classmates and to critically appraise a friend's work. But, as we suggest above, sharing test items with colleagues for review can enhance your assessments. As a pre-service teacher, how might you begin to incorporate such a practice into your coursework? Could you begin writing sample test items for content in your classes and share them with your colleagues?

Summary

Selected-response assessments are useful classroom assessments for a number of reasons.

 • They provide quick feedback about strengths or areas of improvement in student under-standing.

 • Compared to constructed-response items, they are relatively easy to grade.

 • They allow for reasonable comparisons among students or classes.

 • They result in scores that can be easily summarized and presented to parents or administrators.

 • They prepare students for the context and format of the types of standardized testing that they will encounter throughout their academic careers.

Well-constructed test items have characteristics specific to the type of item.

 • True-false items should examine an important idea rather than a trivial fact.

 • Answers to true-false items must be either true or false.

 • Multiple-choice items can be used to assess thinking at various levels, such as factual knowledge, procedural knowledge, and cause and effect.

 • Matching items are used to assess students' understanding of the relationship between ideas.

 • Matching items should include short, homogeneous lists.

 • Matching items should avoid having the same number of items in each column.

133 134

Key Terms

content validity (112)

criterion validity (112)

distracter (120)

equivalent forms reliability (111)

internal consistency reliability (111)

inter-rater reliability (111)

matching test (113)

multiple-choice test (113)

selected-response items (113)

test-retest reliability (110)

true-false test (113)

For Further Discussion

 1. How would you respond to a teacher who states that she only uses essay tests because selected-response tests only assess surface knowledge or recall?

 2. Student guessing on true-false and multiple-choice items is a persistent criticism of such item types. Number a piece of paper from 1 to 25. Now, next to each number, write down either a, b, c, or d. This represents an answer key to a 25-item quiz. Now turn to a partner and ask him to try to guess the letter you wrote next to each number. How many did your partner answer correctly?

 3. Think about your experiences as a student. Do you prefer selected-response items on assessments that you have taken, or do you prefer constructed responses, such as in-class essays? Why do you prefer this type?

Comprehension Quiz

Below are examples of selected-response items designed for an undergraduate psychology course. Your task is to critique the questions with respect to the criteria presented in this chapter. Read each item and rewrite it so that it conforms to all criteria of a well-written selected-response item. There may be more than one error in each item.

 1. There are two major viewpoints about children's cognitive development embraced by educators today, who are the psychologists?

 a. Vygotsky and Piaget

 b. Piaget and Freud

 c. Vygotsky and Freud

 d. Freud and Erickson

 2. What is temperament?

 a. Is a person behavioral style and characteristics ways of responding.

 b. Is distinctive thoughts.

 c. Is how individuals adapt to the world.

 d. Is problem solving and decision makeing.

 3. Which of the following is not something that early maturing girls are more likely to do than late maturing girls?

 a. Have an eating disorder.

 b. Go through depression.

 c. Hold back in the classroom.

 d. Date

 134 135

 4. True or False: B. F. Skinner is the Russian psychologist who is famous for classical conditioning.

 5. True or False: Everything we know is learned.

References

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. L'Année Psychologique, 12, 191–244.

Marsh, H. W., Ellis, L. A., & Craven, R. G. (2002). How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. Developmental Psychology, 38(3), 376–393.

Plucker, J. “Human Intelligence. Alfred Binet.” From http://www.indiana.edu/~intell/binet/shtml.

Relevant Website Resources

University of Minnesota—Office of Measurement Services

http://oms.umn.edu/oms/index.php

This website provides a broad array of measurement and assessment resources. Under the “Classroom Resources” link on the main page, you will find useful reminders and pointers on the development of multiple-choice and true-false items. Bookmark this page for its user-friendly recommendations for assessment in multiple educational contexts.

University of Oregon—Teaching Effectiveness Program

http://tep.uoregon.edu

Under the “Resources” tab, this useful teachers' website presents sound recommendations for the development of multiple-choice items that assess higher-order critical thinking skills. Specifically, this website succinctly summarizes Bloom's taxonomy of cognitive development and then provides sample items and design considerations. Techniques, such as case study items and incomplete scenario items, are provided as examples of items that assess critical thinking.

CHAPTER 6 Constructed-Response Assessments

Chapter Objectives

After reading and thinking about this chapter, you will be able to:

 • Articulate the advantages and disadvantages of constructed-response items.

 • Describe the essential characteristics and evaluate varied examples of short-answer items and essay questions.

 • Construct items that assess students’ ability to articulate their unique understanding of concepts and relationships.

 • State the considerations teachers must make to ensure that students’ constructed responses are assessed fairly and reliably.

In 2005, the College Board, an organization that publishes and scores many widely used educational assessments, added a third component to the familiar SAT. The new SAT writing test provides students with a topic, and students’ responses are scored by experienced teachers. They assign scores between 1 and 6, with 6 being the highest score. The criteria for grading students’ writing include development of a point of view and evidence of critical thinking, use of appropriate examples and evidence, coherence and progression of ideas, skillful use of language and vocabulary, and proper use of grammar.

Not long after the College Board announced its writing test, some people began asking critical questions: How important is writing style? Does the assessment of this writing task demonstrate reliability? Does a timed test of writing authentically measure the skills that go into effective writing? One of the more humorous responses to the SAT writing addition was an article entitled “Would Shakespeare Get into Swarthmore?” (Katzman, Lutz, & Olson, 2004). This clearly rhetorical question about assessing student writing can lead to deeper questions about how and why we should use constructed-response/essay items as a means of assessing learning.

137 138

In Chapter 5, we identified the relative strengths and advantages of multiple-choice, true-false, and matching items, which provide a means of assessing student learning at a variety of levels. This chapter describes forms of assessment that require students to construct a response in order to demonstrate meaningful understanding. In other words, students use their own words to express their understanding at different levels of complexity. Both short-answer items and essay questions require a constructed response.

Foundational Questions for Your Consideration

 • In what ways might essay questions provide authentic evidence that students have learned?

 • A number of standardized tests (the SAT, for example) now include a writing component. Can such administration and scoring of writing samples adequately capture students’ writing ability?

Constructing Responses to Test Items

In this chapter, we look at constructed-response assessments, items that teachers might present to see whether learners are able to construct a proper response. We first examine short-answer items, and we conclude with essay or extended-response items.

It may seem unorthodox to study short-answer and essay items in the same chapter, but they are similar in several important ways. First, as we have mentioned, although these items can be written to test recall, they do not rely on a student's recognition of a term or word. Second, both short-answer and essay items are open-ended, but to different degrees. If we think of open-endedness as a continuum from very open to very restricted, short-answer items are at the very restricted end. Short-answer items are written with a single, brief, correct answer in mind, and, to be counted correct, the students must restrict their responses to that answer. Essay items, however, can be written in such a way that they fall at all parts of the continuum. At the more restricted end, essay items might ask students to produce a short list in response, while at the unrestricted end they might ask students to produce a unique analysis of a concept or theme.

138 139

Ask Yourself

Various technologies have been developed to assess student writing. So, instead of having a teacher or expert read your essay for content, clarity, organization, and understanding of concepts, you might, during your academic or professional career, have your writing analyzed and scored by a computer that recognizes key words, phrases, and structures in composition. How might the use of such innovation influence the way you or your students write? Do you believe that removing the human perspective from scoring undermines writers’ creativity or divergent thinking?

Short-Answer Items

Short-answer items are assessments that ask students to supply a focused answer using their own constructed response. Such items differ from true-false and multiple-choice questions in that they ask the student to supply an answer rather than to select the correct answer from several possible answers. Also known as fill-in-the-blank items, short-answer items assess primarily factual recall—names, dates, places, or specific persons.

What Are the Advantages and Disadvantages of Short-Answer Items?

Compared to true-false or multiple-choice items, students are less likely to guess correctly on short-answer items. In both true-false and multiple-choice items, the answers are provided for the test taker. That is, the answers may be only true or false or one of several possible answers in a multiple-choice item. Not only are short-answer items less subject to blind guessing, but correct responses cannot be identified by simple recognition, as they can for multiple choice.

On the other hand, short-answer items focus only on recall of information and not on higher-level thinking. They require careful attention to clarity in the stem to avoid ambiguity. The scoring also may be more subjective than with other simple types of assessment items, and for this reason would be more time-consuming.

Format Options for Short-Answer Items

There are two major format options for short-answer questions: (1) completion or fill in the blank and (2) question or command.

Completion or fill-in-the-blank items are constructed of a sentence from which one or more words are missing. A blank line is inserted in the 139140sentence, and the student is to write in the missing words at that point. The first option is illustrated below.

 • Dividing a baseball player's number of base hits by his or her total number of official at-bats yields a statistic known as the player's batting average.

 • Abraham Lincoln was born in the state of Kentucky.

The first example shows a typical formatting in which a definition or description is written, with a blank at the end in which the student is to write the name of the item being described or defined. The second is a simpler item, asking for the name of a state. Note that the stem specifies “state” to assure that students are directed toward the required answer. Suppose that was left out and the item read,

 • Abraham Lincoln was born in \_\_\_\_\_\_\_\_\_\_\_\_.

Now the item is ambiguous, and students will not know what sort of response the teacher has in mind. Is “a log cabin” a correct response? What about “1809”? If you want the students to respond with the name of a state, be sure you have written the stem clearly to ask for it.

Here are some more examples.

 • Written during the reign of King John, the Magna Carta was the document that has had the greatest impact on our modern democracy.

 • The vernal equinox is the point at which the sun appears to cross the celestial equator from south to north.

 • Our earth is a part of the solar system, which also includes seven other planets.

In these examples, the blanks are inserted somewhere within the sentence rather than at the end. Again, it is important to remember when using this formatting option to give a clear and complete statement in the item. Do not simply lift a sentence from the text and insert a blank for a key term. Further, you must include enough of the context to direct the student toward the correct answer and away from irrelevant alternatives.

Because it can be difficult to write a natural-sounding sentence while leaving out a word or two that can be readily recalled by a student who has studied the material, teachers often lean toward the second format option, the question or command. This option tends to be easier to write clearly, as no blank space is used that might add ambiguity. Here are two examples, one written as a question and the other as a command.

 • What are the two prime movers of soil that cause erosion?

 • List the two prime movers of soil that cause erosion.

The format is simple in this option, and the task is clear to the students. In order to be graded as correct, they must list two things, which are well-defined by the stem. At the same time, this format can place some restrictions on the complexity of the thinking that the student will do in order to answer. Compare these two examples:

 • When you divide a baseball player's number of base hits by his or her total number of official at-bats, what statistic do you obtain? (batting average)

 • A baseball player's batting average is calculated by dividing a player's base hits by his or her number of at-bats.

The first item, in the form of a question, only asks the student to recall the name of a baseball statistic after reading its definition. The second item requires a more complete understanding of the statistic, in that the student must supply the two components of the formula that define it.

The preceding examples demonstrate levels of recall or understanding for students who are able to read, but how might you assess a kindergartner who has not developed effective reading skills? After a unit on farm animals, you might read questions aloud to the entire class and ask them to draw pictures instead of writing answers:

 • What kind of farm animal gives us milk?

 • What kind of farm animal gives us eggs?

Understanding can be communicated through various symbols (remember Vygotsky?). Allow students to express their understanding at a level that is consistent with their developmental abilities by adapting a variety of ways of expression.

What Characteristics Make Short-Answer Items Effective?

Effective short-answer items are written clearly and unambiguously. A good place to start in writing an item is with the response you are looking for. What is it that you want students to recall, and in what context should they recall it? Answering these questions will help you determine how best to approach writing the question.

You will also need to consider the ramifications of poorly written items. As we mentioned earlier, to reduce the possibility that students are reading their texts only to memorize the textbook definitions, you should avoid creating short-answer items by copying the textbook definition and then simply leaving out a key word to be filled in by the student.

Another common mistake in writing these items is the unintentional insertion of clues to the question's correct answer. Although such clues often find their way into multiple-choice or true-false items, they are 141142perhaps most prevalent in short-answer items. Take a look at the following short-answer items:

 • Intelligence quotient, or IQ, is a standardized measurement of \_\_\_\_\_\_\_\_\_\_.

 • One of the oldest and most commonly used intelligence tests is known as the \_\_\_\_\_\_\_\_\_-\_\_\_\_\_\_\_\_\_\_\_.

In the first question, the test writer likely copied a textbook definition of IQ and added a space at the end of the statement for student recall. Of course, using the word intelligence in the stem of the question unintentionally supplies the test taker with the correct term. The second question, however, might have been somewhat less obvious except that the test writer inserted a hyphen in the answer blank, suggesting that the answer is compound or hyphenated. Many psychology students could recognize this as a clue that the teacher is looking for the Stanford-Binet as the correct answer.

Another possible clue is the length of the blank. Instead of drawing short blanks for short answers and long blanks for long answers, make all blanks the same size. The Stanford-Binet example above also illustrates this unintended clue, with the first blank obviously longer than the second, suggesting a longer word followed by a shorter word.

The most important factor, though, in creating a good short-answer item is clarity. It is all too easy for ambiguity to surface in your short-answer items. When you write a short-answer item than can be correctly answered in more than one way, you will find yourself arguing with the students who have picked the “wrong” way and seek to justify their often-plausible answers. Here are a few more examples:

 • Psychology is \_\_\_\_\_\_\_\_.

The teacher who wrote this item wanted the students to write a short definition as learned in class (“the science of mind and behavior”). But the lack of direction provided for the students allows other options—some plausible, some not so—for example, “a division of the social sciences,” “a complex field of study,” and so on. A clearer presentation of this same question is the following:

 • The science of mind and behavior is called \_\_\_\_\_\_\_\_\_\_.

Here is another way that ambiguity can arise:

 • The two most desired characteristics in a test are \_\_\_\_\_\_\_\_\_\_and \_\_\_\_\_\_\_\_\_\_.

The teacher writing this item is expecting the responses “reliability” and “validity” and will be dismayed to read “clarity” and “precision.” Here is a better way to write this item:

 • The two most desired characteristics in a test are reliability and \_\_\_\_\_\_\_\_\_\_.

142 143

By providing one of the two important characteristics of a test, the teacher directs the students to recall the other.

When Are Short-Answer Items Useful?

When would you choose short-answer items over other item types in developing a test? You might consider writing a short-answer test when students are just beginning to build a vocabulary in a particular area, such as in an introductory biology course. Or perhaps you would use them as part of a course's pretest, to determine how familiar your students are with the terms they will encounter in your course. Short answers are widely used to get a quick assessment of students’ knowledge (recall) of a recently studied topic.

Ask Yourself

Compare a time that you took a multiple-choice assessment with a time that you took a short-answer assessment. Which type of assessment better represented what you knew about a subject area? Why did one type of assessment provide a better Essay Items

Essay items, as we have suggested, allow students to communicate a unique, constructed answer to a question. The major difference between essay items and short-answer items is that a short-answer item focuses on a highly specific response and greatly limits the degree of student construction. There is little room for individuality, nor is it expected or desired. Essay items allow students more choice in their constructing answer and consequently permit greater individuality in their responses.

There are different types of choices that students can be allowed to make when developing a personalized response. For example, in some essay questions, learners are allowed to select their vocabulary, use their own style of writing, or employ graphics and other visuals to provide a response. Another choice in essay questions is that learners sometimes are permitted to select the precise content or the examples that they wish to use to support an answer. Given these possibilities, an important consideration for you as a teacher when constructing an open-ended item is the degree to which you intend to narrow or broaden your students’ responses.

143 144

Resource for Your Assessment Toolkit Constructing Short-Answer Items

One way of approaching a short-answer item is to think of it as a multiple-choice item that is open-ended and without distracters. Recall that one difficulty in writing multiple-choice items is in developing plausible distracters. In addition to the correct answer, each multiple-choice item must have three or four alternate answers that are either incorrect or less plausible than the correct answer.

Consider the following multiple-choice item:

 • Which of the following U.S. holidays originally celebrated the end of World War I?

 a. Flag Day

 b. Labor Day

 c. Memorial Day

 d. Veterans’ Day

This question meets the criteria for a well-constructed multiple-choice item. All of the information needed to answer the question is contained in the stem of the question, and each of the distracters is a plausible alternative to the correct answer (d). Its limitation, of course, is that, in terms of factual memory, it asks students only to recognize the correct answer. It does not ask them to recall and produce the name of the holiday.

As a short-answer question, a similar item might read:

representation of your knowledge?

 • Which U.S. holiday originally celebrated the end of World War I?

This question now assesses students’ ability not only to recall but also to write the name of the holiday. It does not offer the students any inappropriate clues to help them eliminate possible wrong answers, nor does it allow them to guess the right answer.

Guidelines to Writing Good Short-Answer Items

 1. Construct the stem so that the answer is definite and brief.

 2. Make sure that there is only one correct answer.

 3. Avoid lifting sentences from the students’ textbook.

 4. For completion and fill-in-the-blank formatting:

 • Make the response blanks equal length.

 • Avoid grammatical clues preceding the blank.

 • Do not use too many blanks in one item—usually no more than two.

 • Include enough information in the stem to ensure the desired response.

The Continuum of Restrictiveness of Constructed-Response Items

The category of test items covered in this chapter can run the gamut from a tightly restricted response to a broad and unrestricted response, depending on what you want to assess. We have examined short-answer items at the restricted end of the continuum. Essay items are by nature broader than short-answer items. But essay items themselves can also be structured along a continuum from more restricted response to more extended response.

144 145

At the restricted-response end of the essay continuum are essay questions that intentionally limit students’ answers. While these questions rely more heavily on recall of information, rather like short-answer items, they go beyond such items by requiring students to recall and organize information, to structure a personal response, or to briefly defend a conclusion. The following are examples of restricted-response items:

 • List each of Howard Gardner's intelligences and describe a brief (1–2 sentence) classroom activity you might use to develop each of the intelligences.

 • Identify each of Piaget's developmental stages, and in a sentence or two, explain the cognitive abilities that children display during each stage.

 • List three advantages and three disadvantages of using multiple-choice items in classroom assessments.

 • What are the three branches of the U.S. government? For each branch, write a 1–2 sentence summary of its function as defined in the U.S. Constitution.

Each of the above examples demonstrates students’ ability to recall important concepts, such as each of Gardner's multiple intelligences. But the task for students goes beyond recall because they are asked to apply their understanding to the use of multiple intelligences in the classroom. These items go beyond a simple one- or two-word response and require more time to score, but they are still easier to evaluate than the more open-ended essay questions. They also lie somewhere between short-answer items and broader essay items in the extent to which they allow students to construct unique responses that give more insight into their thinking and learning.

As essay questions become more open-ended, they can be considered extended-response items. At the extended-response end of the continuum, we are not limiting students’ responses and are intending to assess higher-order outcomes such as evaluation, organization, analytical reasoning, or originality. Extended-response items, however, do not simply ask students to relate everything they know or can recall about a topic. Instead, the question or prompt itself directs students toward the type of information or skill the teacher is trying to assess. Consider the following:

The Continuum of Restrictiveness of Constructed-Response Items

The category of test items covered in this chapter can run the gamut from a tightly restricted response to a broad and unrestricted response, depending on what you want to assess. We have examined short-answer items at the restricted end of the continuum. Essay items are by nature broader than short-answer items. But essay items themselves can also be structured along a continuum from more restricted response to more extended response.

144 145

At the restricted-response end of the essay continuum are essay questions that intentionally limit students’ answers. While these questions rely more heavily on recall of information, rather like short-answer items, they go beyond such items by requiring students to recall and organize information, to structure a personal response, or to briefly defend a conclusion. The following are examples of restricted-response items:

 • List each of Howard Gardner's intelligences and describe a brief (1–2 sentence) classroom activity you might use to develop each of the intelligences.

 • Identify each of Piaget's developmental stages, and in a sentence or two, explain the cognitive abilities that children display during each stage.

 • List three advantages and three disadvantages of using multiple-choice items in classroom assessments.

 • What are the three branches of the U.S. government? For each branch, write a 1–2 sentence summary of its function as defined in the U.S. Constitution.

Each of the above examples demonstrates students’ ability to recall important concepts, such as each of Gardner's multiple intelligences. But the task for students goes beyond recall because they are asked to apply their understanding to the use of multiple intelligences in the classroom. These items go beyond a simple one- or two-word response and require more time to score, but they are still easier to evaluate than the more open-ended essay questions. They also lie somewhere between short-answer items and broader essay items in the extent to which they allow students to construct unique responses that give more insight into their thinking and learning.

As essay questions become more open-ended, they can be considered extended-response items. At the extended-response end of the continuum, we are not limiting students’ responses and are intending to assess higher-order outcomes such as evaluation, organization, analytical reasoning, or originality. Extended-response items, however, do not simply ask students to relate everything they know or can recall about a topic. Instead, the question or prompt itself directs students toward the type of information or skill the teacher is trying to assess. Consider the following:

These two questions about the life cycle of the darkling beetle ask for specific and narrow information. In other words, the teacher is looking for evidence that the student knows and can articulate the life cycle—egg, larva (the infamous mealworm), pupa, adult (the infamous beetle). There are four distinct stages to the cycle. Students may describe or identify the stages differently (for example, they may identify the larval stage as the mealworm 146147stage), but the question calls for no elaboration and expects an answer within a narrow range of vocabulary.

Multiply Focused Constructed-Response Questions

This type of essay question allows students not only to use their own wording but also to choose their own example, adding another layer of individuality to the response. In the singly focused items, the teacher expects students to identify and describe the characteristics of the life cycle of a particular organism. In the multiply focused question, however, the teacher expects students to reveal a conceptual understanding of life cycles—not just the life cycle of the darkling beetle. The multiply focused item, then, allows students to articulate their understanding of a life cycle in their own words rather than in the anticipated words of the teacher.

 • List the underlying characteristics of all life cycles. Then select one example of a life cycle and describe the characteristics that make it a life cycle.

In this example, students must first abstract and describe the factors that define life cycle. Beyond understanding how darkling beetles move through four separate stages, students are expected to demonstrate what life cycles look like in other species. Then they can use the darkling beetle life cycle, if that is the one they have selected, to show the specific stages that make it a life cycle.

Multiply Focused and Extended Constructed-Response Questions

This type of question allows students to show their understanding of a concept, use their own wording, make some choices about how they will approach the response, and then elaborate on their understanding in a way that demonstrates how they think about something. As teachers, we expect students to demonstrate creativity in their thinking by developing an idea or understanding that is new. But, at the same time, students can demonstrate creativity and unique understanding by elaborating on important concepts or ideas as we demonstrate below.

Multiply focused and extended constructed-response items allow students to demonstrate their unique understanding in several ways:

 1. Showing their work. “Show your work” is a common refrain from teachers, but why exactly are we so insistent on it? From a foundational perspective, the logic that students use to arrive at a solution to a problem is just as important as the answer itself.

 2. Connections to real life. Students’ understanding of an idea or concept may often be more clearly articulated if they are allowed to use real-life issues or examples.

 147 148

 3. Connections across disciplines. When students articulate how a key principle or concept from one discipline applies to another (say, from math to physics), they demonstrate an integration of knowledge that is a form of higher-order cognition.

Questions like these are helpful for both teachers and students. They enable students to clarify their approaches to problems, and they provide insight to the teacher concerning the way a student understands a concept. Consider the example shown in Figure 6.1 of an extended-response essay question (drawn from mathematics) that asks students to show their work.

Figure 6.1 Example of a “Show Your Work” Essay Item.

Calculate the area for the following figure. Show how you went about solving this problem.

Requiring students to show their work gives the teacher insight into the approaches that students use to solve a problem concerning the area of an irregular geometric figure. You will be able to see different levels of sophistication in students’ thinking and you will also be able to find errors in their thinking. The errors students reveal here will be especially valuable to you in improving your teaching of this type of problem. Figure 6.2 shows an essay question that connects the students to a real-life situation and asks them to solve a geometry problem and interpret the answer.

Think about what these different types of essay questions are asking students to perform. Notice how the different questions provide students with more freedom to respond in unique ways. Each question requires students to go beyond the task of simply calculating area or perimeter. 148149Rather, these questions ask learners to explain a principle of geometry, not to just solve the problem.

Figure 6.2 Example of a “Connections to Real Life” Essay Item.

You have been asked by a friend to help build a fenced area for her new dog Gunnar. She has 64 meters of fencing and wants it to provide as large a usable space as possible.

The two rectangles shown below are the two possibilities allowed by the current house and its lot. Which of the two rectangles would you recommend to your friend? What would you say to your friend to explain the reasons why that rectangle is a better choice than the other?

A third type of extended-response essay question allows students to support an answer by making connections drawn from other theories, disciplines of study, or research findings that they have uncovered themselves. Such a question fits an interdisciplinary instructional approach (often employed in social studies) in which students are explicitly taught to analyze arguments from different disciplines of study. Here is an example of an extended-response essay question that builds on students’ interdisciplinary instruction:

 • Drawing from at least two perspectives (e.g., sociological, economic, philosophic, etc.), analyze the causes of the American Civil War.

In this extended-response question, the learners are free to select and make connections drawn from various disciplines. There is a requirement that they employ at least two distinct disciplinary perspectives, but they are free to select the perspectives they wish, depending on their unique interests and backgrounds of study.

In conclusion, note that there are many different ways to allow students to elaborate and provide an extended response. The key for you as the teacher is to be clear in the stem of the question about the content and thinking that is required and the type of constructed response that students are allowed to provide. The examples and the categories we have demonstrated will help you to determine the degree of openness or restriction appropriate to the learning outcomes you need to assess.

What Are the Advantages of Essay/Constructed-Response Items?

Essay questions are most effective at assessing complex learning and higher-order skills. When well written, they require students to organize, integrate, and apply their knowledge in a unique way. They ask student to evaluate ideas, provide arguments and justifications, and analyze outcomes. It is important to keep this in mind as teachers develop essay items, because reducing an essay item to a task of simple factual recall in an extended format is a poor use of this item type.

Essays can enhance students’ study habits. Test taking is a skill for which students must prepare and rehearse. Just as teachers sometimes develop objective test items by pulling a definition straight from the text and changing a word or two, students often use a similar habit in studying—memorizing bold terms and developing mnemonic aids to remember them. Effective essays, however, require that students demonstrate a deeper conceptual understanding by analyzing information or by organizing information in new ways. If students know that their assessment will include essay questions, they are prompted to go beyond simply memorizing a few key points from a chapter.

149 150

Digging Deeper What Might Vygotsky Say?

From a psychological foundations perspective, how do constructed-response items effectively capture student learning? Lev Vygotsky and other social constructivist psychologists argue that allowing students to construct and convey understanding in their own words is the clearest evidence that a student has learned something. For example, if we assess students’ ability to define key terms in a chapter, a word-for-word presentation of the definition does not necessarily demonstrate learning or understanding. But if we observe one student explaining the concept to another student in a clear, coherent, and unique way, we have more direct insight into our student's understanding. It is important, Vygotsky would say, for us to observe the many and varied ways that our students communicate understanding—ways that go beyond paper and pencil.

Essay questions are relatively easy and time efficient to construct. Of course, you would not want ease of development to be the major rationale for using a particular item type, but well-developed and purposeful questions that are clearly linked to learning targets can be written and refined in much less time than a multiple-choice test, for example.

What Are the Disadvantages of Essay/Constructed-Response Items?

Scoring essays is quite time-consuming. Students are interested in the grade attached to the essay test, but to respond to students with meaningful comments about the content—as well as style, grammar, or other elements of the essay—can demand as much time as your students took to compose their responses.

Because essay responses are often fairly long, essay tests give a deeper but narrower picture of a student's mastery of content. This mastery of the material is in comparison to a 30-item multiple-choice test, which may take the same amount of time to administer to your students.

Subjectivity in scoring can be a problem. This is a question of reliability, or consistency of assessment results (remember that reliability refers to the likelihood of a similar score on a particular assessment administered to the same student over time). Unlike an effective and well-written multiple-choice item, for example, students and teachers are certainly aware of the possibility of two teachers scoring an essay differently.

Unintentional bias in scoring is possible. Expectations about student work, and even teacher mood, may influence grades.

These last two points are important considerations to address as you prepare to grade your students’ papers. We will respond to these points later in the chapter, when we discuss scoring.

150 151

Guidelines for Meaningful Essay Items

 1. Identify your purpose. One of the most important guidelines to keep in mind is the appropriateness of essay questions to the concept or outcome you are trying to test. What are the learning outcomes that you are assessing? Essay questions are only appropriate if you want to measure complex learning outcomes and higher-order skills. If your learning outcome is one for which you can use objective questions, then it is inappropriate for an essay question. Once you have chosen essay questions as appropriate, your specific purpose will direct you to the type of essay item to prepare, either a restricted-response item or a more open-ended and extended-response item.

The following stems illustrate learning outcomes that assess higher-order thinking skills, such as analyzing or critiquing, that are appropriate to essay questions.

 • Compare the following two methods …

 • Present arguments for and against …

 • Analyze the strengths and weaknesses of …

 • Critique the following argument in favor of …

 • Using evidence from your reading, explain why you agree or disagree with …

 • Describe a situation that illustrates the principle of …

 • Describe and analyze the factors that …

 2. Be specific and thorough. Use the stem of the question to define the problem completely and identify the specific skills to be demonstrated. Use as much detail in the question stem as necessary. You will find that it is not a waste of time for your students to read longer directions because this will help them write a to-the-point answer. A good essay question specifies how students are to present their understanding, rather than leaving them to write everything they know about the topic. How might good and poor questions differ? Consider the examples in Figure 6.3.

 Figure 6.3 Problems Often Found in Essay Questions with Suggested Revisions

 3. Prepare to score the responses. Write your scoring plan or rubric at the same time that you write the question. This takes time, but, by having your scoring criteria clearly in mind, you will write a clearer question, and your students will be more likely to write meaningful answers. We will discuss how to develop your scoring guide in the next section.

 4. Give every student the same test. Avoid offering students a choice about which essay item they address. When students respond to different 151152 152153questions, they are in essence taking different tests. This is a concern for two reasons. First, since your assessments are linked to the learning targets that you believe to be important, you will want to know how each of your students has developed his or her understanding of those targets. Second, students will most likely respond to the items that they feel most comfortable and confident in answering, and you will not know how well students understand the material in the questions they chose not to answer. That content is equally important or you would not be writing an essay question about it.

 5. Time the test properly for your students. As part of the test instructions, indicate the approximate time to be spent on each question. This will help the students to pace themselves. And be sure to allow for thinking time. You are expecting your students to construct a thoughtful and meaningful essay, so be sure to allow them time to think through their response.

Ask Yourself

In the first meeting of your history class, your professor tells you that he will administer paper-and-pencil assessments every Friday, but he will be alternating formats. One week he will present a 20-point objective test, and the next week he will administer an assessment of two short essay questions. As you leave class that first day, do you think that the different test formats would cause you to study differently? How might the type of assessment influence your preparation for the weekly test?

Scoring Essays

Your class has finished its hour-long, in-class, constructed-response assessment, and you sit down at your desk to review their responses. It has been a long week, and as you scan the papers, you notice that some of the handwriting is sloppy, almost illegible. Several students have obvious spelling and grammatical errors throughout. The length of responses varies from two to five handwritten pages. How will these factors influence your assessment of the students’ responses to your carefully crafted questions?

The Issue of Subjectivity and the Halo Effect

Students and teachers alike should be aware of the subjectivity of scoring essays and essay items on tests. Your students should, of course, recognize that the grade or score that an essay receives represents your appraisal of their work relative to the criteria that you have established and presented 153154to the students through your learning targets and your expectations for the particular assessment. And you, as the teacher, should make every effort to ensure that your assessment of student essays is as objective as possible and not arbitrary or related to your knowledge of individual students.

What can you do to ensure that your assessment of written work is reliable and objective? First, it helps to consider a factor that has been shown to undermine teachers’ ability to evaluate essays fairly and reliably: the halo effect.

We have had extensive experience in working with gifted and high-achieving high school students. One of the characteristics of many such students is that they demonstrate tremendous verbal facility at a young age. By the time these students have reached junior high and are required to write in more complex and analytical forms, their papers are often easily distinguished from their classmates’ by the rich and varied vocabulary. Use of such a vocabulary, however, does not guarantee that the students have mastered the skills of developing their ideas, marshaling their evidence, setting forth a meaningful argument, and so on. But their papers often sound so much better than their peers’ papers that they receive high grades simply on the basis of the vocabulary used. This is an example of the halo effect, meaning that an irrelevant factor can act as a “halo” around the essay, making it appear better than it really is.

The opposite can happen as well. If teachers know the identity of an essay's author, their attitude toward the student may be reflected in comments to the student or even in the final grade. When a teacher dislikes a student for any reason—say inappropriate classroom behavior—the teacher may find himself grading that student's essay more harshly than it deserves.

We offer two suggestions to offset possible bias in scoring:

 1. Establish criteria for grading that focus specifically on the skills that students are supposed to be demonstrating in their writing. Communicate specifically beforehand what elements of students’ responses will be taken into account in the assessment of their work.

 2. If possible, try not to identify the student when reading an essay response. This can be difficult, especially when you read a series of drafts from each student before they hand in a final version of a paper. But every effort that you make will be a step toward a more objective scoring of student-constructed responses.

Developing an Effective Scoring Guide

In order to rate essay responses appropriately and fairly, it is necessary to develop an objective system of scoring in a situation that, by its nature, is subjective. How can we assess the elements of students’ written work and 154155still allow them the freedom and creativity that distinguish effective verbal communication? The answer is a scoring guide (sometimes called a rubric).

Because it is critical that both you and your students understand what will constitute an effective and appropriate essay, your scoring guide must give students a clear set of expectations for their responses. The scoring guide does this by identifying a number of key elements in the expected answer and specifying how the instructor will assign numerical ratings to students’ written answers. (Scoring guides or rubrics may also be developed for assessment of other forms of student performances or demonstration of processes. In this chapter, we focus on scoring guides for the assessment of written work. We will return to the subject of rubrics for all sorts of assessments in Chapter 9.)

But what constitutes an effective scoring guide? At a minimum, scoring guides list the key components in the essay that will be graded as well as the levels of performance that will receive points. A short description defines each level of performance along with the number of points that level will receive.

Figure 6.4 is an example of a basic scoring guide designed to score essay responses to an extended-response item at the high school level. This scoring guide identifies four elements that will be assessed: development of a thesis, the use of appropriate language, the presentation of appropriate sources, and conventional spelling. Each of these elements is assessed along a 4-point scale. Such a scoring guide answers two questions: What specific elements or characteristics in this essay will be assessed, and by what criteria will those same elements be assessed? It is important to note to students, for example, that you will be assessing them on spelling. It is equally important to note that you will not be assessing their formatting of citations, length of paper, or grammar.

Figure 6.4 Example of a Scoring Guide for a High School Level Extended-Response Item

For this essay, the teacher will be assessing the student on four elements, but you can see how this scoring guide might be used again in a slightly different context. Perhaps in a subsequent essay in the same history course, you would include these same elements and add elements that assess students’ ability to compare causes of other conflicts, examine the advances in military technology, or properly format their paper for publication.

What is most important, though, is that you communicate to your students exactly what you will be assessing in their written work. Discuss the purpose behind each component of the scoring guide and be certain that they understand the way in which it will be applied. Not only should you share your scoring guide with students beforehand, as you administer the assignment, but you should invite suggestions about how the scoring guide might further be developed or refined.

Now let's look at Figure 6.5, which is a scoring guide for an essay question at the early elementary level. This essay question might follow a unit on community helpers (such as police, firefighters, librarians, mayors, 155156mail carriers, and so on), in which young students learned the title and function of these people in their community. This simple scoring guide focuses only on the presence or absence of the four elements, rather than on rating the elements.

Figure 6.5 Example of a Scoring Guide for an Early Elementary Level Extended-Response Item

Developing a scoring guide that assesses clearly articulated learning targets and that presents objective and balanced criteria will demand a significant investment of time prior to administering the test, but there are several ideas to bear in mind. First, there will likely be a number of elements that will be common across several assessments, so the scoring guide, 156157or rubric, once developed, will be used in several contexts. For example, if you teach a junior year American literature course, you may assess students’ spelling, grammar, syntax, and citations on each essay assessment. For this reason, you will not have to re-create those elements of a scoring guide for each assessment. Many teachers have a generic starting scoring guide or rubric that they modify, adapt, and add to from assessment to assessment.

It is also helpful to collaborate on the development of a scoring guide or rubric with colleagues in the same grade or discipline. This can enhance the quality and objectivity of your assessments, particularly if there are common assessments in your classes.

Other Considerations in Scoring Essays

In addition to creating a good scoring guide or rubric and preparing yourself to be objective by avoiding haloing factors, consider the following suggestions to help you to fairly assess constructed responses from your students.

 1. Decide how you will deal with factors that are irrelevant to your scoring guide. For example, the scoring guide in Figure 6.4 includes points for spelling but not for grammar and usage. Poor grammar may be one of those factors that could lead you to downgrade a student's essay, even though it is not part of your scoring guide. It is helpful to think through such factors—another might be poor handwriting—to prepare yourself to be objective despite these potential issues in a given student's essay.

 157 158

 2. In tests that contain more than one essay item, score all of one question at one time. This strategy helps to maintain reliability in grading. If you grade all of one question before going on to the next one, you are more likely to follow your scoring guide carefully and evaluate each response from the same framework without being distracted by responses to other items.

 3. In tests with more than one essay item, shuffle the papers after grading all of one item and before going on to the next one. This is another way of avoiding a potential halo effect, either positive or negative. A student's excellent answer to one question could affect your reading of that student's next response, for example. Or you might find that a particular response, which would sound acceptable when read in isolation, sounds either much better or worse when read following a poor or an outstanding answer. By shuffling the papers before reading the responses to the next question, you reduce this possibility for bias in scoring.

While we often think of tests as being either “objective” or “essay” format, it might be useful to consider how an understanding of essay—or constructed-response—tasks can inform other types of classroom assessment. Over the course of your career as a teacher, you will construct many and varied paper-and-pencil tests, comprising true-false, multiple-choice, matching, short-answer, and essay items. Well-constructed items of any format provide a particular insight into your students’ understanding and give you a sense of how they have approached your learning targets. And in scoring objective items, you will arrive at a relatively precise score or percentage correct, because for each of the items, there is one right or best response.

But what about essay tests? Is there a right answer or a proper way for students to respond? What does a numeric or percentage score on a student's essay test communicate about the student's understanding? And what about reliability—a common and legitimate concern of students? That is, how is it that two teachers might score the same essay quite differently?

The reason that we do not refer to essay items as objective assessments is that, when grading students’ written work, we always bring a certain amount of our own perspective to the process, so it is by nature a more subjective form of assessment. But your task as a teacher is to develop an objective means of assessing student work that reflects students’ unique construction of knowledge and communication of ideas.

In a typical course on assessment, a common homework assignment is for students to develop a scoring method and attempt to score a sample essay. This is valuable practice for all, regardless of the age level or subject you will teach, because all teachers will be faced at some point with the need to assess student activities that do not have simple and obvious right answers. For example, how do music teachers assess original compositions? How should P.E. teachers assess swimmers? How can we fairly 158159appraise a student's performance in a class play? As you consider strategies for evaluating essays, note that essay questions represent one example of a larger class of assessments that require special attention to the scoring method. We will address this larger class in the next three chapters.

Resource for Your Assessment Toolkit Advice from a History Professor

Take a look around you the next time a professor returns a term paper or an essay. How many students turn directly to the last page of the paper to find out their grade and then turn to classmates to compare grades? Now, how many of those students do you think will consider deeply how much time the professor has spent reading and commenting directly on student work?

Our colleague—a historian and fine writer—recommends a highly effective way of allowing students to understand the assessment of their written work and improve their skills as writers. At the beginning of each semester he tells his students that if they intend to study history, they should be prepared to think and write like historians, who use a specialized form of communication (as do specialists in other areas, such as psychology, law, or education). Our colleague provides detailed writing assignments, both in his in-class essay exams and in his midterm and final research papers. He clearly articulates the learning targets, identifies specific considerations for each paper, provides clear purposeful questions, and shares his scoring guide or rubric with his students when distributing assignments for papers.

When he assesses the students’ papers, he thoroughly reads each paper and comments throughout using his distinctive fountain pen. In this first reading, he comments only on those elements of the paper that directly influence students’ recorded grades—proper citations, support from primary sources, grammar, and the like. Then, he reads each paper a second time and comments only in pencil. This second round of comments is more rhetorical in nature and leads students to be better writers of history; these comments do not influence the grade. For example, “Would a historian state this as you do?” Or, “You might want to read Livy's original text on this subject.”

Do all of his students go on to become historians? Certainly not, but by the time they finish his course, they recognize that writing is a skill that requires deliberate attention to the ways in which ideas can be communicated.

Ask Yourself

Students often ask whether any consideration of “effort” should be cal-culated into the final grade of an essay or research project. Can we as teachers ever fully discern how much effort a student has put into his or her paper? Think about the example of the gifted students. It is tempting to infer that a student who has provided good examples, used rich, 159160descriptive language, and provided authoritative references in her paper has demonstrated significant effort. Can we as teachers be sure that she worked harder on her paper than the student whose vocabulary on the topic seems limited, whose references are from questionable sources, and whose arguments are weak and misinformed?

Summary

 • Constructed-response assessments include both short-answer and essay items. They are intended to offer insights into student understanding that are not as readily gained using objective type items.

 • Well-developed constructed-response items have several advantages:

 • They allow teachers to assess higher-order skills and complex learning outcomes.

 • They allow students to articulate their understanding of an idea in their own unique way.

 • They do not rely simply on factual recall (dates, places, events, definitions, and so on).

 • They enhance students’ study habits.

 • For teachers, they are relatively easy to construct and link to specific learning targets.

 • Essay items are subject to several disadvantages, in particular:

 • Scoring reliability is a problem for essay tests. In other words, how can we make sure that two readers of the same constructed response would grade it the same way? If scoring of essays is not done in a planned, objective manner, we run into the problem of poor reliability.

 • We need to eliminate such subjective influences as knowing the identity of the writer or the mood of the teacher during grading.

 • As teachers we need to ensure that students are responding to questions that present specific tasks and that are linked to clear learning targets.

Key Terms

completion item (139)

constructed-response assessment (138)

essay items (143)

fill-in-the-blank items (139)

halo effect (154)

multiply focused and extended constructed-response items (147)

question or command format (140)

rubric or scoring guide (155)

short-answer items (139)

singly focused constructed-response items (146)

For Further Discussion

 1. Essay questions and short-answer questions allow students to communicate their understanding in ways that are not entirely prescribed by the teacher. How might such 160161assessments effectively be used in courses such as mathematics? Music? Art?

 2. How might you be able to rephrase the following questions so that they are appropriate essay items?

 a. Discuss global warming.

 b. Should gambling be legal?

 c. What caused the Second World War?

Comprehension Quiz

Part One

Here are three examples of constructed-response items designed for different courses. Your task is to critique the questions with respect to the criteria presented in this chapter. Read each item and rewrite it so that it conforms to the criteria of a well-written essay item.

 1. Discuss Vygotsky's theory of intellectual development.

 2. Evaluate Darwin's theory of evolution.

 3. Was the United States justified in entering the conflict in Vietnam?

Part Two

Choose one of your rewritten questions and devise a basic scoring guide that would assess what you believe to be the appropriate elements on which to assess students’ responses.

Relevant Website Resources

The College Board

In this chapter we reference several times the writing assessment that has become part of the SAT. This website will provide you with an overview of the writing component, details of how the writing component is scored, and sample prompts.

http://www.collegeboard.com/student/testing/sat/about/sat/writing.html

ACT

ACT also offers an optional writing test. Like the College Board Site, the ACT site provides similar information and offers students strategies for test taking and sample student essays.

http://www.actstudent.org/writing/index.html

Brigham Young University testing handbooks

http://testing.byu.edu/info/handbooks/WritingEffectiveEssayQuestions.pdf

This online handbook has the subtitle, “A Self-Directed Workbook for Educators.” It includes information on advantages and disadvantages of essay questions, guidelines for writing good questions, and a self-check set of review exercises.

Essay Scoring Manual for the Regents’ Exam at Georgia Tech University

http://www.lcc.gatech.edu/regents/scoremanual.html

This site, while it addresses the assessment of standardized essay tests, provides useful resources for understanding the reliability issues associated with scoring essay items. It also gives examples of model essays and sample questions.

Reference

Katzman, J., Lutz, A., & Olson, E. (2004). Would Shakespeare get into Swarthmore? Atlantic Monthly, 293, 2.

CHAPTER 7 Assessment Through Observation and Interview

Chapter Objectives

After reading and thinking about this chapter, you will be able to:

 • Discuss the teacher's role in student observation.

 • Identify observable behaviors and skills that match standards and desired outcomes.

 • Construct observational checklists and rating scales.

 • Develop checklists for students to self-assess their work.

 • Describe how to use an anecdotal record or a student interview in assessment.

If ten runners are competing in a mile-long run, and you stand at the finish line and record the times of the runners as they cross the finish line, you can draw several conclusions—the order of finish, the relative difference in time of completion for all of the runners, the margin of victory for the winner, and so on. But what do we not know from simply observing the runners as they cross the finish line? Had we carefully observed runners from start to finish, we might have a different perspective on our race. Maybe one runner was slightly hobbled by a nagging sprained ankle, and another runner lost a shoe in the last quarter mile. Or perhaps two runners collided, and one dropped out of the race altogether. How much richer will our understanding be if we observe and consider the circumstances of the runners and their performance throughout the race?

Teaching is driven by what you observe in your classroom. You note when students comprehend a topic quickly and require less direct instruction or when they have little or no prior knowledge and need instruction at the most basic level. You notice students behaving on and off task. Through simple observations, you can tell when students follow directions to complete an assigned project and when they do not.

163 164

In this chapter, we describe how to use your skills as an observer to reliably and validly assess your students. We will consider the kinds of skills and tasks that can be assessed by observation in formal and informal settings. We present procedures for individual and group observations and offer steps for creating anecdotal records, observation checklists, rating scales, interviews, and student self-assessments.

Foundational Questions for Your Consideration

 • Can observations be entirely objective?

 • In a sense, aren't all assessments observations? What distinguishes observational assessment from other assessments?

 • How do you document observations without introducing your interpretation? Is it acceptable for teachers to interpret behaviors?

Observation: A Direct Means for Getting to Know Students

Observation is a direct means for learning about students, including what they do or do not know and can or cannot do. Students simply exhibit their natural behaviors as they move through the school day in the classroom, the lunchroom, and the playground. This type of assessment closely matches the metaphor of an assessor sitting next to the student. As described in Chapter 1, this metaphor relates to the Latin root assidere (meaning to “sit beside”) and implies that assessment is a natural process of getting to know students by listening to them and observing their behavior. Observations involve an observer who notices when specific student actions or behaviors occur or when they do not. Researchers from Jean Piaget and Lev Vygotsky in the past to Thomas Guskey (2000), Robert Marzano (2000), Arthur Costa and Bena Kallick (2004), and Richard Stiggins (2004) today have promoted observation as one important method for examining how students think and learn.

Gathering observational information makes it possible for you to plan ways to encourage students’ strengths and to improve their weaknesses. You can use data gathered through observation to make decisions about how best to differentiate teaching methods and motivational strategies. For example, sometimes we notice attention-grabbing situations such as students fighting or defying authority. In these conditions, teachers monitor 164165behaviors and define skills that need to be learned or practiced. Sometimes we notice students struggling with a particular concept or process. In these situations, teachers try to identify the most effective approaches that will help the particular student understand the concept or apply the strategy. In all of these circumstances, you will find yourself assessing your students by watching them and then taking appropriate actions based on what you have observed.

Several teachers were discussing the non-progress of a third-grade En-glish language learner (ELL) student in reading comprehension. None of her written comprehension tests met grade-level expectations. The enrichment teacher had observed that the student had outstanding drawing skills and suggested that the student be allowed to answer comprehension questions through illustrations. Her reading teacher took her aside for private questioning/interview and encouraged her to answer the comprehension questions in pictures. Her visual responses showed that she did recognize the beginning, middle, and end of the story; she could identify the major characters and the setting of the story; and she could clearly distinguish the hero from the villain.

The enrichment teacher's informal observation, combined with a more formal interview, uncovered some important information about this ELL student that had not been revealed in other, more typical, assessments. A student who appeared to be failing was shown to meet all of the expected goals for the lesson when she was assessed in an individualized, adapted situation.

Advantages of Assessment Through Observation

Immediacy

Observation allows us to assess our students as we are teaching. We can monitor progress and behavioral skills as part of the normal teaching process. Observation of skills and knowledge can take place in the natural teaching and learning setting of the classroom without the need for a specific test or assignment. For example, the physical education teacher watches students on the ball field as they play a baseball game. Even though students are intent on winning the game, the teacher gathers important insights about their abilities to hit or field the ball. Or the math teacher observes as students in a group create graphs on the computer to represent the data they collected for their research project. Both examples allow the teachers to observe students applying their knowledge of specific skills in lifelike, natural situations without the intervention of a specific test.

Unique Information

Observation allows you to discover skills and detect problems that would be difficult to uncover in any other way. For example, a student who tests well on knowledge-based questions may show gaps in learning 165166when placed in a natural context that requires the student to retrieve or apply that information. That gap in learning can be detected and then corrected by using observation in the natural context where a test would be inappropriate.

Differentiation

Observational methods can give us clues that permit us to adapt other assessments to student needs, as in the case of the ELL student. The teacher's informal observation provided insight into a different way for that student to show what she could do.

Value Added

Observation can add a missing dimension to our assessment of students. Information that you gain from observing a student can be used together with other, more formal, assessment methods, such as paper-and-pencil testing.

Disadvantages of Assessment Through Observation

Subjectivity

There are several hurdles for a teacher to overcome when observing students, because this form of assessment can be subjective. We may form a faulty judgment based on a single instance of observation or from another teacher's report. We may set expectations that are unrealistic for our students. We can also be biased for or against a student due to prior positive or inappropriate interactions.

Time Factors

Another hurdle for the classroom teacher is the time needed to conduct observations. You cannot do the work of observing your students from your home, as you do when you bring home a briefcase full of essays to grade. Observation notes taken in the classroom or elsewhere need to be recorded as soon as possible. If you observe a student on Monday and wait until Friday to complete an observation form or make notes, you are compromising the accuracy of the assessment.

Also, if you audiotape or videotape the student's behavior or performance in order to share with others, you will have to spend time replaying the tapes and recording the data. You will have a record of the student's behavior as documentation, but the act of taping and then replaying the tape to take notes is time-consuming.

Uncovering the Right Behaviors

Other difficulties in observing students lie in the reliability and validity of the observation. If you do not specify the explicit skills to be observed and carefully plan your observations, you may end up observing unrelated behavior. For example, talking loudly is more noticeable than staying on task or being productive. You can become distracted from your target observation by extraneous behaviors or other irrelevant factors if you have not developed a clear and careful plan for the 166167observation. A written checklist targeting the explicit skills or behaviors will help the observer focus on the purpose of the observation.

Digging Deeper Binet and Simon

Two French psychologists, Alfred Binet and Theodore Simon, are credited with the development of one of the first widely used intelligence tests for children. The Binet-Simon scale involved an assortment of tasks that Binet and Simon believed to be characteristics of children's abilities at different ages. In an effort to confirm their hypotheses about these tasks, both psychologists used extensive observation studies focused on children of different ages. Employing a variety of observation tools and examining data drawn from these tools, Binet and Simon developed a set of tasks ordered by difficulty and complexity level.

Once the tasks were developed, Binet and Simon tested them with 50 children—10 children each in five age groups. Teachers had identified these children as possessing average intelligence for their ages. Each child was then asked to tackle the Binet-Simon tasks ranging from simple to increased complexity. The test items ranged from very simple (shaking hands, pointing to parts of the body) to slightly harder (repeating three digits or simple sentences spoken by the examiner) to more challenging items (creating sentences from several words given by the examiners, describing differences between objects). “The hardest test items asked children to repeat back seven random digits, find three rhymes for the French word obéisance, and to answer questions such as ‘My neighbor has been receiving strange visitors. He has received in turn a doctor, a lawyer, and then a priest. What is taking place?’ (Fancher, 1985)”

Adaptations of the Binet-Simon scale continued to be used after Binet's death. The most famous, the Stanford-Binet scale, was adapted and named by Lewis Terman of Stanford University and covered both children and adults. The most recent version of the Stanford-Binet scale is used by psychologists to determine the aptitude of children and adults.

It is noteworthy that the work of Binet, Simon, Terman, and their successors is based on a careful series of observations. Observation remains a powerful assessment tool in that it provides a direct and often unobtrusive means for determining what children know and can do.

Source: Fancher (1985) quoted at “Human Intelligence”: Alfred Binet, http://www.indiana.edu/~intell/binet.shtml). Accessed January 14, 2008.

Identifying Observable Behaviors That Match Standards and Desired Outcomes

Clearly, knowing what to look for is the key to sound observation. Are you observing a desirable behavior that you want to increase, or are you targeting an undesirable behavior that may be inappropriate or even harmful? Academic skills, psychomotor skills, and prosocial skills are behaviors that we want to increase or improve in our students. Obviously, inappropriate 167168behaviors are those we want to decrease. Both appropriate and inappropriate behaviors can be the focus of an observation.

Academic Skills

All of the academic subjects are included within the category of academic skills: reading, mathematics, science, social studies, language arts, foreign language. Each assessed cognitive skill area must be stated in observable, measurable terms. Moreover, the vocabulary for the observable and measurable terms should include the knowledge, skills, and concepts listed in each state's academic standards and benchmarks.

In general, you would develop a tool to observe academic skills that are displayed as part of some performance, like reading aloud, recording data, calibrating a balance, using technology or science equipment, participating in a dialogue in another language, or giving an oral report.

Psychomotor Skills

The category of psychomotor skills includes gross motor and fine motor skills, physical movement as in various sports, dance, and physical exercise; the performing arts—singing, playing an instrument—as well as art and craft production, such as painting, drawing, and building a model. A flutist observed playing a specified piece of music might be assessed on pitch, tempo, tone, posture, and breath control. If one or more of these component skills is observed to be lacking, the teacher can target those areas for remediation.

Prosocial or Affective Skills

Included here are the skills students need to work and interact with others appropriately. Prosocial skills involve student's attitudes, beliefs, feelings, or the dispositions they develop. These skills must be taught to students just as purposefully and precisely as academic skills.

Often the specific and more observable social skills are derived from the principles that set the climate for the class. Classroom principles are usually guiding precepts such as strive for excellence, respect yourself and the rights of others, accept responsibility for your actions, participate in your own learning, and cooperate with others.

It is difficult to assess these lofty goals without breaking them down into specific behaviors that students will understand. Figure 7.1 lists some observable social skills that are written in “kid language” so students understand the expectations.

Figure 7.1 Prosocial Learning Behaviors

We often place students in groups and tell them to “work together,” “cooperate,” or “be a team.” We assume students know how to collaborate, and, when we do not get the behaviors we expect, we sometimes abandon cooperative group work and return to individual seat work in frustration. By breaking down the prosocial skills needed for individual or group work, selected skills can be taught and reinforced through the regular curriculum, 168169allowing you to observe your students and diagnose problems that the students may have, as well as offer interventions to increase achievement and prevent harm.

Adapting Your Vocabulary to Your State Standards

Each state has adopted its own standards and benchmarks (indicators, descriptors, elements, and proficiencies) that are assessed on the state's high-stakes standardized tests. Therefore, it is critically important that you create your academic skills checklists with the end in mind—and the end is the vocabulary, concepts, and principles of your state standards.

Another challenge involves implementing state standards while at the same time not allowing your textbook to dictate the vocabulary of knowledge and skills. The vocabulary in the textbooks may be different from the vocabulary used in the state standards and on the standardized tests that are aligned to those standards. Therefore, it is important to include the vocabulary from both sources—the textbook and the standards—in your instruction and assessments so students do not become confused. Figure 7.2 compares an example of textbook vocabulary to the vocabulary of a particular state's standards.

Figure 7.2 Textbook Versus State Standards Vocabulary

Building Reliability and Validity into Your Observations

Reliability

In the context of observation, reliability refers to consistency and to agreement among observers. Suppose you and another teacher independently 169170observed the same student at the same time, using the same method, and then compared notes. Would you come up with the same results? Would you agree on what you saw or would you disagree? You can see that reliability is closely connected to the issue of objectivity that we discussed earlier. Observers will be more likely to agree on what they see in an observation if they have objectively defined the behaviors they are observing.

Reliability in observation is also associated with immediacy of recording. Observations that are not recorded as you conduct them run the risk of being inaccurate or unsound. How much can you rely on notes that were taken after the observation, when time has passed and you have perhaps forgotten some important details?

In addition, reliability in observation is associated with manageability. If you are trying to observe a long list of behaviors and a number of students, you can easily find yourself overwhelmed by the experience. When there are many things to look for, it is easy to overlook one thing or another. In fact, you will miss some of what you intended to observe if you are trying to do too much.

170 171

Observation Validity

Observation validity refers to accuracy and completeness. Are you really collecting information on the behavior or skill that you targeted? This is the accuracy component. Did you set out to observe cooperation in a group of students and then discover that your results relate to the one student who did all the talking during the group work? Defining behaviors and skills is not easy, as we discussed earlier. “Cooperation” is not an observable behavior but an inference drawn from the behavior in which you see your students engaged. You will need to determine what behaviors you can observe that will be indicators of cooperation. The specific behaviors you select will depend on the group's task, but they might include helping others when asked, saying supportive things to others, and taking turns. In this context you would probably also include such inappropriate behaviors as interrupting others, talking loudly, and refusing to share material. You would be collecting information on a series of behaviors that together could be a definition of cooperation in a particular situation.

Completeness in observation has to do with collecting a representative sample. When you observe your students, you must be sure that you have provided them with a reasonable opportunity to exhibit whatever it is you are observing. Otherwise, if they do not exhibit the skill, how will you know if it is because they cannot or because they did not have the chance? And one observation is never enough. If a student fails to perform the skill you are watching for, is it because you happened to observe her on the one occasion when she did not perform the skill? The only way to know is to observe more than once. This gives you a chance to get a better sense of the level at which the student is currently performing: Is this an emergent skill? Is it progressing? Or is the student showing mastery? These questions are answered through multiple observations.

How Do You Make Your Observations Reliable and Valid?

First, be sure that you have carefully defined the behavior or skill you are going to observe and that you have a clear plan for when and under what circumstances you will observe. As a part of this process, you will plan the type of observation to conduct and will develop a form or checklist to use to collect your data. We will have a great deal to say later in this chapter about the various kinds of observations you can conduct and the formats that can be used.

Second, assure that your observations are reliable and valid by practicing observing your students and your planned behaviors and skills. Following the practice, you can revise your definitions and methods as needed. Often you will find that you have omitted a part of the behavior or skill that will be needed for an accurate observation or that some part of your planned definition is still unclear. Or you may find that the form you have prepared does not give enough room to collect the needed information or the format is 171172awkward to use. Practicing allows you to correct any problems that come up and prepares you to be accurate and thorough in your observation.

Third, when possible, get expert help. Include another teacher in your observation so that the two of you can compare your findings. This gives you another opportunity to clarify your definitions or methods and a chance to get an outside opinion on any subjectivity that might be clouding your observation. Because it can be difficult for another teacher to set aside time for this, you may decide to save this step for the more significant or difficult observations—those that may have an especially large impact on a student or those where you find yourself struggling to define a behavior. And observations that are conducted across your entire grade level will require participation by your other grade-level teachers in this step.

Fourth, give sufficient opportunity for the observed students to show you the skill or behavior you are observing. This is an important issue that you must think through when you are planning the details of when and where to observe.

Fifth, record your observations immediately. This will happen automatically if you have prepared a specific form for the observation and have practiced with it so you can use it easily.

Sixth, review your data after the observation. Be sure that you have filled in all the spaces and notes as you planned and that it is readable so that at a later time you can interpret the data you collected.

You can create and use the tools yourself for your observations or develop them in your grade-level team and arrive at a group consensus of what behavior merits a particular interpretation or rating. The reliability of observation is improved when all teachers using the tools are trained and supervised during a practice session.

Ask Yourself

Consider a student who often disrupts class, talks back, and causes the teacher to feel distressed. How might this inappropriate interaction affect the teacher's ability to fairly assess the student's participation in a co-operative group-work setting?

Systematic Approaches to Classroom Observation

Throughout the school day, teachers regularly observe students and make countless decisions concerning how to respond to students, what to teach next, and what behavior to encourage or discourage. This is part of daily 172173classroom life. Some observations, however, can easily be forgotten in the midst of a busy day. Patterns of conduct may be overlooked because there are simply too many distractions and demands.

Resource for Your Assessment Toolkit Observing Inappropriate Behaviors

A critical prerequisite of successful observation is to carefully determine (1) what behaviors you are looking for and (2) what specific questions you are trying to answer concerning those behaviors. For example, if you notice that a student exhibits off-task behaviors during sustained silent reading time, you could develop some specific questions that might guide some focused observations. Your objective for these additional observations might be to determine if the student has trouble staying on task in a variety of situations, in certain specific situations, or only during sustained silent reading.

Points to consider when observing inappropriate behavior include the following:

Setting: What was the situation in which the inappropriate behavior occurred? What was happening just before the student's behavior?

Behavior characteristics: What exactly did the student do? Was she off task by playing with a ruler at her desk, or was she off task by running around the room, knocking over other students’ books?

Frequency: How many times was the child off task during 20 minutes of silent reading?

Length of time: How long did each off-task behavior or episode of the behavior last?

Latency: How long did it take before the student followed directions, took out a book, and began silent reading?

Consequences: What was the effect of the undesired behavior? For the off-task student who runs around the room, pushing books off desks may be reinforced by her classmates’ laughter. This makes the behavior much more difficult to modify, and the teacher may need a completely new approach that involves the entire class, not just the off-task student.

Observation tools are instruments and techniques that help teachers to focus and to record useful data about students’ learning in a systematic way. The observation tools we will discuss are anecdotal notes, observation checklists for teachers, student checklists for self-assessment, rating scales, and interview guides. Each of these tools will help you record important observations. Together they provide a powerful source of assessment data that flows directly from the natural rhythm of the classroom.

Anecdotal Notes or Records

The least formal type of observation uses anecdotal notes or records as a means to document observations of significant skills, events, or behaviors of students. Anecdotal records are a quick, open-ended way to record observations, describing what happened in the teacher's own words. They 173174are intended to record factual descriptions of meaningful incidents that the teacher has personally observed. Their purpose is to enable teachers to quickly note a behavior or concern or noteworthy event so that at the end of the school day, they can reflect on and better interpret patterns of conduct that need attention or deserve recognition.

Impromptu Anecdotal Records

Even though you can plan in advance to use them, anecdotal records are often used spontaneously for unusual classroom incidents. You might find it most helpful to limit this kind of recording to capture instances of misbehavior or to focus on those students who need your help getting organized or completing academic tasks. Most likely, you will anecdotally record only exceptionally good or exceptionally troubling instances of student skills or behaviors, rather than every student in each skill or behavior. Your records can be used when writing report card comments or in parent or student conferences. They are also useful if another intervention, such as acceleration or specific subject tutoring, appears to be needed for the student.

Figure 7.3 is an example of an impromptu anecdotal record. It describes a behavioral incident and offers the teacher a record of her actions. The record notes only the facts of the incident, without an emotional accounting or evaluation.

Figure 7.3 Anecdotal Notes for Gary M.

The anecdote recorded in Figure 7.3 was spontaneous, but you can plan for such impromptu recording by designing a generic form to use whenever an unusual situation arises in the classroom. You might put the form on an index card, which would be easy to keep nearby and relatively unobtrusive to use. The card might have a few preplanned blanks (student's name, date, setting) followed by an area in which you can record whatever is needed. This format is designed to be nonspecific so that it can be used in a variety of situations.

There is no one right way of keeping anecdotal records. Some teachers keep an informal log in a notebook with information about all students. This is helpful when jotting down incidents after whole-class instruction. 174175Or you might record anecdotal observations as running accounts, one for each student, on separate pages in a notebook. Some teachers use “sticky notes” and paste the notes into files in each child's portfolio. Another option is to use the computer to print general or subject-specific forms on adhesive labels. Then you can observe and record on the label, peel it off, and quickly attach it to a student's folder.

Figure 7.4 illustrates a more detailed anecdotal recording form to be used in an impromptu observation. Because it gathers information about what happened just before and after an incident, it can be used when the teacher suspects that an inappropriate behavior will require some kind of intervention. This is also called A-B-C recording because it includes the antecedent, the behavior, and the consequence.

Figure 7.4 Anecdotal Record (A-B-C Recording) of Inappropriate Behavior in Context

The A-B-C recording provides a way to develop an understanding of a child's challenging behavior (Carr et al., 1994; O'Neill et al., 1997; Hieneman et al., 1999). The record allows the teacher to record certain unusual behaviors within their exact context. The form has a place to record the setting (context) of the event, what happened just before the behavior (antecedent), what the behavior looks like (behavior), and what happens after the challenging behavior occurs (consequences). The more detailed anecdotal record helps the teacher begin to uncover possible causes of the student's behavior and to predict what conditions could cause future behavior in order to develop a potentially useful intervention to decrease its frequency.

175 176

Planned Anecdotal Records

Anecdotal recording can also be used in a planned and systematic way by designing a tool that focuses on specific skills that are part of a larger task like problem solving or taking tests. You can use such a form in each subject area and for a variety of purposes. Figure 7.5 is an example of a skill-specific anecdotal record, focused on test-taking skills. In this case, the teacher is focusing on a student who is performing poorly on written classroom algebra quizzes. This is a student who seems to understand the concepts in class work and discussion but who scores poorly on quizzes. Could the problem be related to test-taking skills?

Figure 7.5 Skill-Specific Anecdotal Record—Test Taking

Note that in this anecdotal record the number of behaviors to be observed is limited so the teacher can avoid distraction, focus clearly, and use the time efficiently. Further, being clear on the overall skill and the behaviors you want to observe with planned anecdotal records will help to make this process more thorough and unbiased. Using this simple form, the observation can be recorded as the student is taking the test.

In any case, it is critical that you record your observation as soon as possible to maintain observation validity. Trying to recall the details of an observation and record them several days later will not result in valid data. Therefore, develop a method that accomplishes your purpose and is easy for you to use on the spot or immediately afterward.

Making Anecdotal Records Reliable and Valid

By their nature, anecdotal records represent an informal method of observation. There are a few simple 176177steps you should take to make your anecdotal records reliable and valid. If you are making planned observations, use the methods we discussed earlier to increase reliability and validity:

 • Define the behavior.

 • Create a form for recording.

 • Practice.

If you are gathering data spontaneously, there are two critical factors:

 • Record the incident as soon as possible.

 • Keep it factual by avoiding interpretation or emotions.

For these spontaneous, or impromptu, records, having a generic form to use can help you get the observation down on paper without delay.

Observation Checklists

Observation checklists are clear and concise lists of behaviors that are used to assess a student's skills. The teacher observes the skills in a simple and straightforward manner, marking them as present or absent, correct or incorrect. Checklists are flexible options that can be used by the students themselves as well as the teacher.

Using a Checklist to Observe Academic Skills

Observation checklists can focus on the processes or tasks needed to demonstrate knowledge, abilities, or mastery of a skill. Each task must be written in a way that is observable and measurable. Such checklists can be used at specific intervals to ensure that a student moves forward with no gaps in his or her learning. Each component of the specific task is listed in order on the checklist.

The first step in preparing to observe an academic skill is to identify the key dimensions of the overall skill. For example, fluency is a common assessment focus in reading: A teacher observes as a student reads aloud and rates the student on fluency. But what exactly is a fluent reader? Fluency is made up of the separate skills of accuracy, decoding, phrasing, and expression. A student may be able to read accurately, decode new words, leave proper phrasing between sentences and characters, but still may read in a flat, expressionless tone. The teacher now has a skill area to target for that student in building full fluency.

For the skill of developing an orally presented argument, the dimensions can include stating a clear position, providing evidence to support the position, using precise vocabulary, exhibiting an awareness of counterarguments, and responding to critics who would not agree with the position. For the skill of presenting the oral argument, the dimensions can 177178include speaking clearly, pausing to allow the audience sufficient time to process the information, and using inflections. These dimensions of the overall skill of developing and presenting an oral argument need to be considered carefully and translated into specific behaviors.

Next, you will need to refine the precise behaviors that relate to the key dimensions in light of the students’ grade level. As noted above there are many different dimensions that relate to the skill of developing and presenting an oral argument. However, the precise behaviors that relate to these dimensions differ based on the maturity of the speaker. For example, when using precise vocabulary, you would expect that first-grade students’ words choices would differ from sixth-grade students’ choices. One way to specify this might be to expect first-graders’ words to be drawn from level one of the Dolch Basic Word List.

A useful step in developing a thorough observation checklist is to identify common errors that relate to the skill. For example, as you observe your students informally, you might note that many students in the class forget to provide evidence that is based on empirical research. If this is the case, you would make certain to specify the presentation of empirical evidence as a specific behavior that deserves special attention and clarity in your checklist.

Figure 7.6 shows a checklist that focuses on the oral presentation skills that we have been discussing. Observations that result from the use of Figure 7.6 are useful for two reasons. First, the behaviors closely relate to a standard that has been selected as especially important. Second, the behaviors are specifically related to known weaknesses of students. By collecting data on these carefully selected behaviors, you can provide specific insights that will help students improve potential weaknesses.

Figure 7.6 Middle School Oral Presentation Observation

How do you translate the information from the Figure 7.7 checklist into a meaningful description of proficiency? One way to use the same checklist to take account of growth over time is illustrated in Figure 7.8. It includes a proficiency criterion and is designed for use across multiple observations throughout the school year. Columns have been added to the left-hand side of the checklist to allow you to make observations at three points during the year. You then derive a score on each skill by summarizing. You would circle the YES (proficiency is achieved) if the student has displayed the skill more than half of the time (in this case at least two out of three times) or circle the NO if the student has not. The date at the top of the form would be the point at which you summarized your three sets of checkmarks.

Figure 7.8 First-Grade Locomotion Skills Observation Checklist, End-of-Year Summary

178 179 179 180 180 181

Using a Checklist to Observe Prosocial Behavior

Checklists can easily be constructed to observe students as they work alone or in group settings. The steps for creating a checklist for observed behaviors are similar to creating an academic or psychomotor skill assessment checklist.

First, be sure to define the behaviors in observable, measurable terms on the checklist. “Listens to group members’ ideas” or “Offers positive feedback to team members” are desired behaviors for a group interaction. “Raises hand when asking questions” may be a desirable behavior for an individual observation.

Second, establish a manageable period for observing. For example, if you are observing students as they work in groups, observe one group a day. If you are observing individual students, plan to monitor no more than four or five per day.

Third, give your students a copy of the checklist and explain why and when you will be using it. Make sure to inform students on the day you are observing them. And give them opportunities to grow comfortable with being observed.

Fourth, give the students feedback on what you have observed. Reinforce positive behavior and provide scaffolding and modeling for students who show harmful, inappropriate, or undesirable actions.

Figure 7.9 is a checklist that can be used when observing a cooperative group. Note that the behaviors are described explicitly in observable terms. The teacher should inform the group they are being assessed on their cooperative skills and place herself in close proximity to the group so she can watch all students. The students are given a check when the behavior is observed or NA when the behavior is not apparent. There is a place for comments or anecdotal notes. The group should be observed more than one time, especially if any of the group members did not have the opportunity to demonstrate all the skills. In fact, you may need to focus on one student at a time until that student has had sufficient opportunity to display each skill and only then focus on another student. No matter what approach you use, the key is to observe all students long enough to provide a valid observation. Information from the checklist should be shared with the group members after the observation.

Figure 7.9 Observation Checklist for Cooperative Group Work

Observation checklists provide a roadmap that helps teachers give students consistent feedback related to the task. But in addition to teachers using checklists to gather information about students, it is important to include students in the process so they can observe and self-assess their academic progress. This shift in focus from the teacher as third-person objective observer to the student as first-person stakeholder in the process can foster metacognition and self-assessment. These checklists are valuable tools in our assessment repertoire.

Figure 7.10 is a checklist designed for observing an individual student's behaviors. Observing one individual at a time is time-consuming for the 181182teacher, but it can offer specific information about a student's work and study habits. Furthermore, the cooperative group-work checklist in Figure 7.10 could be used by the group members themselves to rate their behavior and participation.

Figure 7.10 Observation Checklist for Cooperative Group Work

Using a Checklist to Assess Dispositions

Dispositions, that is, patterns of behavior that are valued by our society, are included in all of the national standards. For example, the national science standards state that perseverance 182183 183184is an important disposition to nurture. By placing learners in a situation that requires their perseverance (such as making daily observations on a science project and taking notes), the teacher can collect evidence by noting the care and constancy of the daily notes that learners keep. This evidence provides data to assess perseverance.

The summary checklist in Figure 7.11 is based on multiple observations and focuses on four dispositions: work habits, study habits, persistence, and social skills. It can be used by the teacher, a peer, or the student, and it has a space for comments and future goals. Sometimes making the students aware of the behaviors to be observed will help them become more conscious of demonstrating those specific behaviors.

Figure 7.11 Checklist for One Student's Behaviors and Dispositions

Making Your Observations Reliable and Valid

In order to make your observations as reliable and valid as possible, consider the following suggestions.

First, think carefully about what it is that you are trying to observe so you can answer these questions:

 • What behaviors will I look for?

 • What is an appropriate number of behaviors to observe that will tell me what I need to know about the students without overwhelming me?

 • What activity or context will give the students the best opportunity to show me those behaviors?

Second, having developed your definitions and formatted your checklist, practice observing several times to make sure your definitions hold up and your checklist is easy to use. Third, focus carefully when conducting the observations to avoid distractions. Fourth, when reviewing your data after the observation, be sure that you have completed everything in the checklist that you intended to and that you can clearly interpret what you wrote at a later time.

Rating Scales

A rating scale is another tool used to observe student skills and behaviors, in addition to anecdotal records and checklists. Rating scales are actually a form of checklist. In this case, the checklist has been modified by adding descriptive words, numbers, or both to rate the adequacy or appropriateness of a student's behavior. Typically a rating scale consists of a list of qualities that are judged according to a scale that indicates the degree to which each quality is present. The key that distinguishes the rating scale from other checklists is the assumption that each characteristic on the scale 184185 185186can be observed according to some underlying degree of accomplishment. Therefore, “more” accomplishment means the person will be rated higher on the rating scale.

Descriptive Rating Scales

The basic rating scale is descriptive and is based on a series of adjectives or thumbnail sketches. The descriptions form a rising scale intended to describe different degrees of progress toward a desired learning outcome. They allow the teacher to rate the adequacy or appropriateness of a student's behavior along that scale. The “best,” or most desired point, is last on the scale, and the other points along the scale are thumbnail sketches of how students’ performances look along the continuum.

As with all forms of observation, you will construct a rating scale by first specifying the observable behaviors that are important in a particular context. Perhaps you have a short list of skills that a student is struggling with in a particular academic area. Having selected and defined those skills or behaviors, you will need to construct a rating scale. An easy method for writing the adjectives that will describe the points on the scale is to determine the best and worst likely performances and then choose in-between levels to create the full scale. Making the description as specific as possible enhances accuracy during the rating process.

Figure 7.12 displays a rating scale designed to be used when ­observing a student working math problems individually. It focuses on three math skills: working on problems, checking work, and correcting ­mistakes—generalized skills that could be ­contributing to the student's problems staying on task. This teacher has selected three levels or rating points for use in this rating scale, but the specific points differ for each skill. By creating specific points that are unique to each skill, the teacher has zeroed in as clearly as possible on the behaviors being observed. The result will be detailed information that can be used directly to remediate any problems that are uncovered during the ­observation.

Figure 7.12 Descriptive Rating Scale of Mathematic Skills

Numerical Rating Scales

The numerical rating scale associates numbers with descriptions along the scale. In most cases the higher the number the greater the accomplishment, and lower numbers imply lower accomplishment. Numerical scales are often used when summarizing observations across some period of time. You might use this type of scale after observing a series of discussions over time or after observing student behavior in a classroom over several weeks. For example, the numbered points within a rating scale could be based on the number of times a particular behavior has been noted. This kind of scale then could look as follows: 1 = never; 2 = 186187sometimes; 3 = usually; 4 = always. Clearly defined descriptors enhance the rating process as in the example here:

1 = Never. Behavior is not observed.

2 = Occasionally. Behavior has been performed, but repeated instances of nonperformances are observed.

3 = Usually. Behavior is performed, but a small number of instances of nonperformance are observed.

4 = Always. Behavior is consistently and regularly performed.

You should avoid using rating scales with numbers only—that is, without descriptive adjectives associated with each number—because without the adjectives there is no indication as to what differentiates a 3 rating from a 4 rating. It is sometimes recommended to have an even number of points to avoid clustering ratings in the middle.

You should keep one important caution in mind when using numerical rating scales. The numbers in the scale do not necessarily represent equal intervals. That is, it may not be the same distance in terms of skill development if a student moves from a rating of 1 to a rating of 2 as it would be to move from a 2 to a 3 or from a 3 to a 4. The numbers are used in a rating scale almost as if they were a different kind of descriptive adjective 187188themselves. At the same time, the advantage of using numbers in the scale lies in the ease with which we think about and understand numbers. We are all thoroughly accustomed to recognizing that 4 is larger than 3, 3 is larger than 2, and so on. If the descriptive adjectives used in a particular rating scale are in any way unclear, the numbers remind us instantly which category is which and which direction is “up.”

Figure 7.13 is a numerical rating scale for cooperative group work. As students work on the project, the observation focuses on the steps of group dynamics leading to the final project. When the entire class is working on a project or problem in cooperative groups, it is best to target only two or three groups at a time.

Figure 7.13 Rating Scale for Cooperative Group Projects

You can reuse this rating scale every time students work on projects cooperatively, keeping track of group members as groups rearrange. You will probably find that different mixes of students will show different observed behaviors. As in all group activities, the downside to this type of rating is that not all members may display the same behavior, yet you score the group as a whole.

Figure 7.14 notes how one student is observed performing in a group. This eliminates the problem of assessing the group as a whole. This type 188189of single-student observation is time-consuming but provides more detailed information.

Figure 7.14 Rating Scale for Evaluating an Individual Group Member

Cautions Regarding Interpretation of Observational Data

A basic feature of student observations is relying on personal opinion. Your conclusions are subject to all the characteristic errors of human judgment. By being alert to the following list of errors, you can avoid inadvertently making them during an observation:

 • Scoring too leniently or too severely.

 • Avoiding the extremes of the scale and scoring at the average.

 • Allowing an outstanding or lesser feature of one performance to influence the scoring of other factors.

 • Scoring a student lower than average if the previous student observation was outstanding or scoring a student higher than average if the previous student observation was not successful.

 • Judging the student according to a personal stereotype or strongly held attitude.

 189 190

 • Judging the student by an initial impression rather than on the basis of multiple observed performances.

 • Rating a student more favorably if the student is similar to you in respect to background, attitudes, or ethnicity.

 • Forgetting observations if not systematically recorded.

Ask Yourself

Consider what it felt like when you were asked to perform a task in front of another person. Perhaps you had to take a driving test with an examiner sitting next to you. Perhaps you were asked to play a musical instrument in front of judges. Or perhaps you had an opportunity to audition for a play or some other performance. Did this observed experience enable you to perform better or did it cause you to make more errors? What does your experience tell you to do when you wish to observe students perform a task?

Interviews

Interviewing is an interaction in which the teacher presents a student with a planned sequence of questions, listens to the responses, asks further questions, and records data. Often interviews are associated with a particular problem with which the student is struggling. These exchanges can be an important source of information about how the student is forming concepts or using procedures.

Robert Marzano (2000) describes the informal interview as an opportunity for the teacher to probe the students’ understanding of a topic in ways that cannot be used effectively with other forms of assessment. He says the informal interview “allows the teacher to pose questions like ‘Tell me a little more about that’ or ‘Explain that to me again but in a different way.’ This form of interaction is potentially the most valid type of assessment a classroom teacher can use” (p. 100). Additionally, interviewing students gives you data for modifying or enriching instruction and curriculum. ­Interviewing can be used with all levels of students, including those who are unable to take paper-and-pencil assessments or who have learning disabilities or disorders. It is especially appropriate with students who appear anxious when speaking in front of the class or who have trouble with writing.

190 191

Resource for Your Assessment Toolkit Self-Assessment Checklist for Sentence-Writing Skills—First Grade

Here is a checklist that the students themselves can use. Even very young students can be encouraged to begin reflecting on their learning and to assess themselves. This kind of checklist will support that kind of reflection.

Interviews provide opportunities to develop rapport with your students as you probe their thinking. As you question a student about how or why she came to a certain conclusion, you learn more about how she thinks, and you can ask follow-up questions to probe more deeply. Besides listening to a student's responses, you have the opportunity to observe him closely. Are there long pauses as he gathers his thoughts? Is he fidgeting in his seat or with a pencil? Does he make eye contact with you? These reactions may give you further insight into a student's abilities and provide you with data that you could not get in other ways.

Using Interviews in Your Classroom

Interviews can be structured or unstructured. In unstructured interviews the teacher asks developmentally appropriate questions that occur naturally in the conversation. These interviews evolve depending on the student's responses to questions. In structured interviews the teacher has prepared the questions in advance. Here are some tips for preparing questions:

 • Word your questions so the students can easily comprehend what type of information is required.

 • Do not deviate from standard classroom vocabulary.

 • Let students know before the interview that they are free to ask for clarification of a question if they do not understand it.

 • Use open-ended questions so the answers are not predetermined, and the students can respond in their own terms. Avoid questions that can be answered yes or no.

Begin all student interviews that have the same purpose by using the same directions and materials. This will help standardize the procedures and validate the data you gather from the interviews, even though you will modify your questions with each individual as you follow up on his or her responses.

Make sure to give the student enough time to formulate a response. Do not rush. If the student appears unable to reply to a question, wait several moments before you move on to the next question. Much of the value of the interview process comes from allowing the student to set the pace.

Decide how you will record the interview. You can use any of the observational methods we have discussed—anecdotal records, checklists, or rating scales—as a data collection tool for your interview. You may only be able to take brief notes during the interview to avoid interrupting the student's flow of thoughts. You can then transfer your notes from 192193the interview to the tool you selected after the interview is completed. If your school permits, you might consider recording or videotaping the session. You will probably only want to record the most critical or difficult interviews, however, as every recording will have to be viewed or listened to again in order to gather your data, doubling the time you spend.

Some students may feel intimidated when you question them individually, but they usually enjoy sharing their experiences with each other. In this case, you may want to consider having students ask each other the questions while you listen to the responses. Students also enjoy sharing with their peers how they solved a problem or got an answer. Questions that ask students to describe how something they read in a story or studied in science or social studies relates to their lives can help students exhibit their learning. When teachers treat student-to-student interviews seriously, a spirit of cooperation and openness develops. Make sure students are comfortable with this arrangement and give classroom practice as needed.

Interviews also allow teachers to address higher-order thinking skills and essential questions that unlock understanding. Since time is limited, interviews could address the higher levels of comprehension, analysis, synthesis, and evaluation. Figure 7.15 gives a list of questions that encourage students to go beyond recall of facts.

Figure 7.15 Interview Questions Focused on a Book Report

Figure 7.16 gives a list of questions for a student interview about mathematics problem-solving skills. After giving the student a specific problem to solve, the teacher would use these questions to allow students to share their abilities, to explain what they are thinking, to clarify their approach, and to display creativity. Further probing questions would accommodate individual differences. This process can be repeated at 193194regular intervals as you introduce new problem-solving strategies in your instruction.

Figure 7.16 Interview Questions Focused on Problem Solving

The interview process can also be used informally with individual students or groups of students to review evaluations from other observation formats. Although interviews can be time-consuming, certain assessment objectives can be better achieved by student interview than by other techniques. Process objectives that focus on understanding the way students solved a problem lend themselves to interviews because the interview setting allows students to use, words, drawings, and manipulatives to display their thinking. As the teacher listens and observes the child, misconceptions and erroneous strategies are uncovered and used to inform future teaching strategies. Moreover, when you have differentiated both your instruction and assessments, students will have selected product and performance options to meet their interests or abilities. Interviews allow teachers to ask specific questions to assess each student's level of understanding related to the tasks.

Processing questions require students to use metacognition as they reflect and adjust their thinking. Reflective questions listed in Figure 7.20 could be used as part of interviews to encourage students to transfer their learning beyond the academic content and make connections between their they learned knowledge and their lives.

Figure 7.20 Reflective Questions for Interviews

In addition, students with limited English skills or weak writing skills may not be able to explain their deep understanding of key concepts in a 195196written essay. Therefore, the interview allows them to express their ideas orally without being penalized for their language or writing weaknesses, organizational challenges, or time management skills. Interviews help balance the assessment opportunities for all students.

196 197

Conducting Reliable, Valid Interviews

The guidelines we have discussed earlier in this chapter are relevant here in making your interviews reliable and valid. Planning carefully, selecting a manageable chunk of material, and immediate recording of answers are all important. In the case of interviews, your planning will focus on the questions you want to ask, of course, but you also need to consider the context that will make the student most comfortable in responding to you. You need to think carefully about the way you set the context to elicit the student's best cooperation. You need to create a data collection tool that can be easily used while interviewing, and you need to practice with the tool you create. And, while interviewing, you need to focus carefully to avoid distractions. Review your notes as soon as possible and add to them where needed to make them as complete as you can. You need to be certain they will be fully understandable when you read them at a later time.

Summary

 • Observation is the process of gaining information by watching and listening to students.

 • Observation techniques can be used to evaluate students’ knowledge, skills, dispositions, and behaviors.

 • Observations based on standards can provide important insights into students’ abilities.

 • It takes time to fairly and equally observe all students and keep records of the data obtained.

 • You can strengthen the reliability and validity of observations by using a clear, concise scoring tool that defines the desired performance criteria in observable, measurable terms.

 • You can align your performance criteria to your state standards.

 • Anecdotal notes, checklists, and rating scales are scoring tools that you can create, modify, and reuse to evaluate specific student skills and behaviors.

 • Teachers can probe a student's thinking through interviews and record the observed data on a scoring tool that can be repeated at specific intervals to show growth.

 • Data gained from observations should be combined with other forms of assessment to gain a clear picture of a student's abilities.

 • Students can use the feedback from checklists and interviews to reflect on their learning and make appropriate adjustments to meet their goals and standards.

197 198

Key Terms

academic skills (168)

anecdotal notes or records (173)

descriptive rating scale (186)

descriptors (187)

interviewing (190)

numerical rating scale (186)

observation (164)

observation checklist (177)

observation validity (171)

rating scale (186)

prosocial skills (168)

psychomotor skills (168)

structured interviews (192)

unstructured interviews (192)

For Further Discussion

 1. What do you see as the most important advantages and disadvantages of using observational methods compared to more traditional paper-and-pencil assessments?

 2. Name three dimensions of an academic or physical skill of your choice and determine the specific behaviors that comproise each skill.

 3. List five prosocial behaviors that would be beneficial for you to record through direct observation. What problems might you encounter in defining these behaviors for observation?

 4. Think of a task that your students perform on a regular basis, such as setting up a lab experiment or writing a journal entry. Create a checklist of all the steps the students need to carry out to complete the task.

 5. How can self-assessment checklists help students take ownership of their learning?

Comprehension Quiz

Each observational scoring tool below has one or more design errors. Identify each flaw and correct it.

 1. Anecdotal Record of Observed Student Behavior

 Student

 \_\_\_\_\_Juanita\_\_\_\_\_

 Date

 \_\_\_\_9/16/2005\_\_\_\_\_\_

 Observer

 \_\_\_Ms. Dayton\_\_\_

 Grade

 \_\_\_6th\_\_\_

 Juanita has not answered any math problems correctly. I am very upset that she doesn't respond to my help. I think she is pretending not to understand me so she can get out of doing her work. I am going to assign her more problems until she understands that she must do her work.

 2. Checklist of Observed Student Reading Skills Fifth-Grade Reading Comprehension Skills

 Name

 \_\_\_\_\_\_\_\_\_\_\_

 Date

 \_\_\_\_\_\_\_\_\_\_\_

 Check YES if the learner can perform the step (✓)

 \_\_\_\_ 1. Likes to read.

 \_\_\_\_ 2. Reads at home.

 \_\_\_\_ 3. Identifies main characters.

 \_\_\_\_ 4. Identifies plot, setting, and conflict.

 \_\_\_\_ 5. Identifies genre.

 \_\_\_\_ 6. Reads for pleasure.

 \_\_\_\_ 7. Summarizes story events.

 \_\_\_\_ 8. Uses neat handwriting.

 8. Rating Scale for Observing

198 199

Rating Scale for Observing Inappropriate Student Behavior

Name

\_\_\_\_\_\_\_\_\_\_\_

Date

\_\_\_\_\_\_\_\_\_\_\_

Relevant Website Resources

Alaska Department of Education and Early Development

http://www.eed.state.ak.us/tls/frameworks/langarts/42tools.htm

The Alaska site includes examples of observation checklists for writing conferences, reading conferences, listening skills, anecdotal notes, six-trait writing, oral presentations, and group projects.

Center for Evidence-Based Practice: Young Children with Challenging Behavior

http://challengingbehavior.fmhi.usf.edu/fba.htm

This website from the University of South Florida contains information about Functional Behavior Assessment, a process that helps develop an understanding of children's misbehavior in specific situations. The process involves collecting information through interviews, using behavior rating scales, and recording setting, antecedents, behaviors, and consequences in order to determine what conditions reliably predict the occurrence of the child's challenging behavior.

Center for Information and Research on Civic Learning and Engagement (CIRCLE)

http://www.civicyouth.org/PopUps/Chi\_checklist.doc

This site includes student observation checklists for civic skills and behaviors such as personal responsibility, caring for the community, leadership, and caring for others. The checklists are arranged by grade levels K–1, 2–3, and 4–5. The checklists are designed to help teachers document student civic development.

National Science Teachers Association

http://www.nsta.org/main/news/stories/science\_and\_children.php?news\_story\_ID=52414

The National Science Teachers Association (NSTA) site includes tools teachers can use to assess students in science. The article “Unlocking the Power of Observation” by Karen L. Anderson, Dean M. Martin, and Ellen E. Faszewski shows how to use checklists to help kindergarten through second-grade learners, particularly students with limited English language skills. It includes an assessment checklist and rubric to 199200assess students’ abilities to use their powers of observation.

TERC

http://www.terc.edu

The TERC website focuses on research-based materials based on national and state standards in mathematics and science. The performance tasks (some are in Spanish) target data literacy, scientific process, problem solving, evidence-based decision making, mathematical fluency, an understanding of the natural world, and student reflections on their own learning.

References

Burke, K. (2005). How to assess authentic learning, 4th ed. Thousand Oaks, CA: Corwin Press/Sage.

Burke, K. (2006). From standards to rubrics in six steps: Tools for assessing student learning, K–8. Thousand Oaks, CA: Corwin Press/Sage.

Burke. K. (2008). What to do with the kid who … Developing cooperation, self-discipline, and responsibility in the classroom, 3rd ed. Thousand Oaks, CA: Corwin Press/Sage.

Carr, E. G., Levin, L., McConnachie, G., Carlson, J. I., Kemp, D. C., & Smith, C. E. (1994). Communication-based interventions for problem behavior: A user's guide for producing behavior change. Baltimore: Paul H. Brookes.

Costa, A. L., & Kallick, B. (2004). Assessment strategies for self-directed learning. Thousand Oaks, CA: Corwin Press/Sage.

Fancher, R. E. (1985). The intelligence men: Makers of the IQ controversy. New York: Norton. Quoted at “Human Intelligence: Alfred Binet,” http://www.indiana.edu/~intell/binet.shtml. Accessed June 5, 2007.

Guskey, T. R. (2000). Evaluating professional development. Thousand Oaks, CA: Corwin Press/Sage.

Hieneman, M., Nolan, M., Presley, J., De Turo, L., Robertson, W., & Dunlap, G. (1999). Facilitator's Guide: Positive behavioral support. Positive Behavioral Support Project, Florida Department of Education.

Johnson, D. W., & Johnson, R. T. (1989). Cooperation and competition: Theory and research. Edina, MN: Interaction Book Company.

Johnson, D. W., & Johnson, R. T. (1995). Teaching students to be peacemakers, 3rd ed. Edina, MN: Interaction Book Company.

Johnson, D. W., Johnson, R. T., & Holubec, E. J. (1993). Cooperation in the classroom, 6th ed. Edina, MN: Interaction Book Company.

Marzano, R. J., (2000). Transforming classroom grading. Alexandria, VA: Association for Supervision and Curriculum Development.

O'Neill, R. E., Horner, R. H., Albin, R. W., Sprague, J, R., Storey, K., & Newton, J. S. (1997). Functional assessment and program development for problem behavior. A practical handbook. Pacific Grove, CA: Brooks/Cole.

Stiggins, R. (2004). New assessment beliefs for a new school mission. Phi Delta Kappan, 86(1), 22–27.