Chapter 12

Assessing for Learning: Objective and Essay Tests

This chapter will help you answer the following questions about your learners:

- How can I make sure that my assessments are fair to students?
- How can I make sure my classroom tests measure what I teach?
- How can I write test questions that require learners to use the same thought processes emphasized during my instruction?
- How do I choose among test item formats for an objective test?
- How do I write objective test questions that accurately measure what I have taught?
- How do I write multiple-choice items that measure higher-order thinking skills?
- How do I write essay questions that accurately measure what I have taught?
- Can essay tests be scored reliably?
- Should I base my grades on how a learner's achievement compares with the achievement of other learners, or should I base them on a standard of mastery that I determine?
- How do I combine different indicators of achievement, such as classwork, homework, quizzes, and projects, into a final grade?

In this chapter youwill also learn the meanings of these terms:

content validity
criterion-referenced grading
extended-response essay
flexible-response tests
grade weighting
norm-referenced grading
response alternatives
restricted-response essay
restricted-response tests
stem
test blueprint
test fairness
test validity

Geri Dalton is principal of Fawkes Middle School. Last summer she attended a workshop on the assessment of classroom learning. As a result of three days of rigorous training in assessment techniques, she decided that her major goal this year will be to improve classroom testing and grading at Fawkes Middle School.

It is now the fall term, just six weeks after the first grades have been sent home. Ms. Dalton has planned individual meetings with her teachers to discuss their testing and grading practices, beginning with her first-year teachers. Richard Steele is first on her list.

Ms. Dalton: Come in, Richard. Please sit down. I reviewed the description of your grading system that you sent home to parents at the start of the year, and I'd like to ask a few questions.

Mr. Steele: Fine. But please be kind. I never had a course in testing in college and it's been trial and error for me.

Ms. Dalton: What about during student teaching?

Mr. Steele: Well, my supervising teacher had a pretty complicated system, but she never really explained it. So what I eventually came up with is a mixture of things I remember from when I was in grade school, practices I saw in college, and whatever I picked up last year from observing other teachers.

Ms. Dalton: That's pretty typical. Most teachers develop a testing and grading system based on the very things that you considered.

Mr. Steele: I'd appreciate any suggestions you have.

Ms. Dalton: OK. Let's first look at what you base a grade on. You're using chapter tests 30 percent, daily quizzes 10 percent, homework 20 percent, notebook 20 percent, and class participation 20 percent. Why did you choose those activities and those percentages?

Mr. Steele: Well, those are the things I have them do in my class, and the percentages are small enough to all add up to 100 percent, yet large enough to be meaningful. That way, if the students don't do the notebook, for example, their grade drops 20 percent. That motivates them.

Ms. Dalton: So tests make up 40 percent of the grade, while homework, a notebook, and participation count 60 percent. You're implying that tests are less important for assessing learning than informal measures. Is that what you want? Do you think parents realize when they see a grade of 90 that most of that was not based on tests?

Mr. Steele: I never looked at it that way. I was just trying to balance everything out.

Ms. Dalton: How do you grade the notebook and participation?

Mr. Steele: Well, it's somewhat subjective. I base each grade on 100 points, and I assign a certain number of points at the end of the six weeks depending on the quality of the notebook and the quality of participation.

Ms. Dalton: I didn't see any criteria for assigning these points in your classroom evaluation plan, so I assume you didn't write anything down. Is that right?

Mr. Steele: Yes.

Ms. Dalton: So the most subjective aspects of your grading system count as much as the most objective—your tests. Is that what you intended?

Mr. Steele: I really didn't look at my system that analytically. I was just doing what I saw other teachers do.

Ms. Dalton: You have a really diverse group of learners in your class. Some are better at writing than speaking, and vice versa. Some have good memories but may be weak in problem solving. Others reason well when you're talking to them but can't do it under the time constraints of an essay test. How do you take into consideration all these diverse learning needs in your grading system? It seems to be based primarily on the written word.

Mr. Steele: I have to give the same kind of test to everyone, don't I?

Ms. Dalton: That depends. Are you assessing to promote learning or assessing to assign a grade?

Mr. Steele: You ask tough questions. I guess to assign a grade, although I really want my tests to improve learning.

Ms. Dalton: Sounds as though your grades may not measure what you value. I guess a more basic question is, Do you teach what you value? Let me make one other observation. Your tests are all essay. Why?

Mr. Steele: I hated multiple-choice and true-false tests in college. All they did was measure good guessing. I'm more interested in how my students think.

Ms. Dalton: That's a good value. But look at your essay questions on your first chapter test. They all ask your students to recall information. Not one question measures thinking or reasoning. Since you're only measuring memory, wouldn't a multiple-choice or true-false or fill-in test be easier to grade? This test had ten questions, and each question required about 50 words to answer. When you multiply that times the 150 students in all your classes, that comes out to 75,000 words that you had to read and comment on. How long did it take you to grade all those tests?

Mr. Steele: Two entire weekends.

Ms. Dalton: If you used a machine-scored answer sheet for multiple-choice questions, it would have taken you about 30 minutes.

Mr. Steele: Do you mean we should just use multiple-choice tests?

Ms. Dalton: No. But if all you're measuring is recall, objective-type tests save a lot of time and your learners get feedback sooner. You've got to be fair not only to your students but also to yourself.

Mr. Steele: I guess there's more to grading than I thought.

Classroom Evaluation Activities: An Overview

Most teachers, from elementary school through college, think of making tests and assigning grades as one of the more bothersome aspects of teaching. They view it as a chore, something they have to do to please administrators and parents, rather than an integral part of teaching.

However, skilled evaluation can have a substantial impact on learners in both the short and the long term (Crooks, 1988). At the lesson and unit level, skilled evaluation can:

- reactivate previously learned skills and knowledge
- focus learner attention
- provide opportunities for learners to practice and consolidate new information
- help learners keep track of their own learning and progress
- give learners a sense of accomplishment.

Over time, skilled evaluation can help learners acquire individual habits of learning that:

- increase their motivation to study
- influence their choice of study patterns and learning strategies
- influence their choice of future activities and courses
- establish a realistic picture of their own abilities and competence.

While most teachers would enthusiastically support these outcomes and acknowledge that evaluation is essential to bringing them about, they also are concerned that their efforts at evaluation may be inadequate (Crooks, 1988).

Teachers use a wide range of evaluative activities. Some are informal, such as questions during a class discussion, written notes on assignments and homework, and oral comments during a practice activity. Others are more formal and systematic, such as teacher-written tests. Significant numbers of teachers have had no formal training in assessing classroom learning, while many of those who have had such

training find it of little use or relevance in evaluating their learning activities (Gullickson & Ellwein, 1985).

Given a lack of teacher enthusiasm for evaluation, it is not surprising that learners feel the same way. Just as some teachers fail to view evaluation as a meaningful requirement for learning, so do many learners view tests and grades as chores not worth the effort.

The unskilled practice of evaluation has many unfortunate results, as the dialogue between Mr. Steele and Ms. Dalton attests. It can lead to assessment techniques and grades that are unfair not only to learners, but also to both their parents and the teacher. It can mask what learners actually achieve during instruction. Some tests, both those made by teachers and those published, bear little relationship to what learners actually do during instruction. This discrepancy is a major reason why some learners fail to prepare for tests. Furthermore, unskilled evaluation practices clearly put certain learners at a disadvantage. Classroom assessment techniques that rely almost exclusively on paper-and-pencil tests may ignore the strengths of some learners in a diverse classroom and provide them little incentive to study and to improve.

Ineffective evaluation practices are unfair not only to learners but also to parents. Most parents are keenly interested in what their children are learning. In many schools, report cards are the only means parents have of determining that their children *are* learning. Depending on your evaluation skills, your grades may or may not tell them what was learned.

Finally, an unskilled pattern of evaluation practice is unfair to you. You will spend countless hours developing, scoring, and grading homework, class assignments, and tests. Your time is one of the most precious commodities you have. The time you spend on assessment should be on activities that accurately indicate the success of your teaching efforts.

Fairness in Assessment

The major goal of this chapter is to help you develop a pattern of evaluation practice that you, your learners, and their parents will perceive to be fair and that is genuinely fair. A fair pattern of assessment is built on the following values. **Test fairness** means that your expert pattern of evaluation practice should:

1. Provide a valid assessment of what you teach.
2. Motivate learners to higher degrees of effort.
3. Be sensitive to differences in gender, culture, and abilities.
4. Communicate performance and progress accurately to learners and their parents as well as to future teachers, admissions officers, and employers.
5. Be efficient, saving you and your learners time and effort.

We will discuss each of these characteristics in the remainder of this chapter. Before we do, let's turn to one other goal of student assessment: the need to be practical and realistic.

While it is important that you teach what you value and measure what you teach, the technology of measurement does not always allow you to do this with scientific precision. Many things you value and teach may be difficult, even impossible, to measure. Furthermore, it is unlikely that you will be able to adjust your tests to individual learners. Inevitably, therefore, some of your tests will be too difficult for some learners and too easy for others, more suited to some of your learners' testing styles than to others, and more motivating for some learners than for others. Be aware that your tests will not be perfect assessments of your learners' performance and progress. Tests, even when well constructed, are only *samples* of your students' behavior that *estimate* their true level of performance and progress. This is why a balanced set of assessment techniques, chosen from those to be described in this and

the following chapter, will be important in establishing an effective pattern of evaluation practice.

Validity in Assessment

> The single most discouraging time during my first year of teaching was after I scored my first test. I put a lot of time and effort into my lessons and I was hoping everyone would do well. I wanted everyone to get an A. No one did, and more than half the class failed. Hadn't they studied? Had I done a poor job of teaching? Was my test bad? They seemed to understand what I was teaching. They did the lab assignments correctly, used the equipment properly, filled out the activity sheets, took notes on what they saw. They answered all my questions during the lessons. I had no idea what went wrong. So I curved all the grades to make everyone happy. That left a bad taste in my mouth, which lingers to this day. (Author, personal experience)

In the fields of psychology and education, most tests that have been developed to appraise a particular mental state or trait, such as intelligence, creativity, attitude, or achievement, do so indirectly. In other words, the behaviors that learners demonstrate on these tests (for example, choosing the correct option on a multiple-choice question, agreeing or disagreeing with attitude statements, identifying whether a statement is true or false) are not of themselves of interest. Rather, they provide an indication—or behavioral sign—of an underlying trait or ability.

Classroom assessment of learning, particularly beyond the early elementary grades, is almost exclusively based on paper-and-pencil tests, which also indicate, rather than directly measure, what children have learned (Gullickson & Ellwein, 1985). For example, we may measure an understanding of the scientific method not by having learners plan, conduct, and evaluate an experiment (a direct measure), but by asking them to list the steps in conducting an experiment, to write about the

difference between a hypothesis and a theory, or to choose the correct definition of a control group from a list of choices (all indirect measures). Or we may measure children's understanding of money not by observing them select food, pay for it, and get the correct change (direct assessment) but by asking them to recall how many pennies there are in a dollar, or to write down how much change they would get back from a $10 bill if they paid $6.75 for a T-shirt (indirect assessment).

There are obvious advantages to indirect assessment of achievement and learning, not the least of which is efficiency. It would be very time-consuming to measure directly all learning that goes on in a classroom. But indirect assessment raises a thorny problem: How do you know that the test is measuring what you say it is? In other words, how do you know it has **test validity**? Recall that *validity* is the degree to which a test measures what it says it measures—and what you want it to measure.

No test is completely valid or invalid. A test is valid or invalid only for a particular purpose—the purpose for which it was built, such as measuring knowledge, self-concept, or problem-solving ability. Without reference to purpose, the concept of test validity would make no sense. A test may be valid or invalid for assigning children to special classes for the gifted (for a test that measures aptitude). Or a test may be valid or invalid for determining which learners are in need of remedial work in a particular subject (for a test that measures prior achievement).

In the previous chapter we learned about the importance of predictive validity in standardized tests. Recall that the *predictive validity* of a test indicates the extent to which a test score anticipates or predicts how an individual will perform at some future time related to what the test is measuring. For example, it indicates how someone might perform in college on the basis of a standardized ability test score received in high school, or how well someone performs on the job after taking a job selection test.

In this chapter we will learn about content validity. **Content validity** is a measure of the degree to which a test covers all the content that was taught in the manner in which it was taught—how well, in other words, the content of your test matches the content of your classroom instruction. In this chapter we will take a close look at content validity.

Ensuring Content Validity

Most teachers respond with "Of course!" when asked if their classroom tests measure what they taught. However, one of the most common complaints of learners, whether in grade school or in college, is that their tests often do not measure what they've been taught (Tuckman, 1988). Very likely, the reason behind the poor test performance of the learners described in the vignette that opened Part IV of this book is that their test lacked content validity. Although the learners may have acquired all the goals and objectives the teacher emphasized, the test may not have measured them. How can this happen?

For a test to have content validity, it must have two qualities.

- It must reflect the goals and objectives of your lessons.
- It must give the *same emphasis* to your goals and objectives as did your lessons.

In other words, content-valid tests ask learners to do what they have learned in class. Not all tests do this. As you will see shortly, tests routinely measure skills different from those intended or taught. Moreover, many classroom tests over- or underemphasize certain content areas compared with the emphasis and amount of time devoted to that content during instruction.

To put it another way, content-valid tests measure what teachers teach and learners learn. They ask learners to do what was modeled, coached, and practiced during instruction. If learners saw a teacher demonstrate how to focus a microscope,

were coached to do this, and practiced doing it, a content-valid test would ask them to focus a microscope—not to label the parts of the microscope on a diagram.

In discussing how to build tests to ensure content validity, we will focus on two types of assessment techniques: **restricted-response tests,** which include true-false, multiple-choice, matching, fill-in, and restricted essays for measuring knowledge, comprehension, and application behaviors, and **flexible-response tests,** which include extended essays, term papers, research reports, and other performance-based assessments of learning that measure higher thought processes such as analysis, synthesis, and decision-making behaviors. In this chapter we will focus on restricted-response tests, and in the next chapter we will turn our attention to flexible response tests representing performance-based assessments. Table 12.1 summarizes behaviors that are best measured by each type of test (Bloom, Englehart, Hill, Furst, & Krathwohl, 1956).

Building Content-Valid Restricted-Response Tests

Learners spend much of their time completing tests or assessments of various types. Evaluation activities such as written, teacher-made tests; standardized tests; classwork; homework; questions embedded in texts; and questions during class occupy a significant percentage of a learner's school day. Merely taking teacher-made paper-and-pencil tests has been estimated to consume an average of 5 to 15 percent of the school day, with the higher percentage being typical of secondary school learners (Haertel, 1986).

Teachers usually fall into three traps when making pencil-and-paper tests:

- They test content areas they didn't teach. This typically occurs when teachers hold learners accountable for chapter content that was not discussed in class, assigned for homework, or encountered in workbook exercises. Often this occurs

when the teacher decides that even though a certain content or skill was not taught, the “good student” should have learned it anyway.

- They place more emphasis on certain content areas on the test than when they were actually teaching these content areas. A common complaint from students is that some or many of the questions on the test covered areas only briefly discussed in class.
- They ask questions in a manner that requires students to use intellectual skills that do not match the way they were taught or the teacher’s intended goals and objectives. For example, they use true-false questions to test whether learners can express relationships, or they use essay questions to test recall.

The Test Blueprint

One way to avoid these testing traps is to construct a **test blueprint.** Table 12.2 shows a teacher’s blueprint for a test in a physical science class. The topics covered are listed down the first column, and the intellectual skills emphasized in class are identified across the top. The percentages indicate the instructional emphasis for both content and intellectual skill. This test is worth 100 points. The totals for rows and columns reflect the emphasis (in total points) that the test must have to match the instructional emphasis provided in class. The cells identify item types that measure the intent of the lesson and the number and point totals of these items.

There is a variety of systems for constructing test blueprints (Kubiszyn & Borich, 1996). Ideally, they should be constructed at the time you plan your lessons and are presenting them to your learners. That way, the content you actually teach and the emphasis placed on it is easily remembered. Using a test blueprint is the best way to assure yourself, your learners, and their parents that your grades reflect what your students have learned from your instruction.

Matching Test Questions and Objectives

The skilled test writer prepares questions that ask the learner to use the same thought processes that were identified in class activities, used in the text, and required by homework. In other words, the test items should ask learners to do exactly what is specified by the objectives that guided your instruction. For example, if you want your learners to recall from memory five major battles of the American Revolution in their correct chronological order, you would ask them, “List five important battles of the American Revolution and arrange them in the order in which they occurred.” You would not ask them whether America won the Battle of Bunker Hill (true-false), since this question would not call for the same mental processes required by knowing all five battles in their chronological order. Similarly, if you want your learners to distinguish a noun from an adjective, you would ask them to compare the words “beauty” and “beautiful,” not to define these two words. Recalling definitions would tell you little about the learner’s ability to distinguish the difference between nouns and adjectives. Furthermore, if you want to assess whether a learner can design and execute a scientific experiment, you might ask the learner to carry out an actual experiment that could be rated for thoroughness, completion, accuracy, objectivity, and so forth, and not simply to list the steps of the scientific method.

When you assemble your test, you will either create test questions or use those provided by your textbook publisher or by other teachers. In each case, the best way to ensure a match between your goals and objectives and your test items is to follow the three steps outlined in the accompanying box, *Matching Test Questions to Instructional Objectives.* If the level of behavior required by the test question matches your objective, you have a content-valid test item. If it doesn’t, rewrite the item to match the objective.

Summary

The expert practice of developing classroom tests requires first that you develop and follow a test blueprint to ensure that the content coverage and emphasis of the questions match your instruction. Second, it requires that you carefully write test items that demand of the learner the same intellectual skills or behaviors that are specified in your goals and objectives. Your decisions about item formats (true-false, fill-in, multiple-choice, essay) will be based largely on which format best helps you measure the desired behavior, thereby achieving content validity.

Choosing Test Item Formats:
General Considerations

In the following sections, we cover specific considerations involved in writing content-valid restricted-response test items (we turn to the techniques involved in flexible-
response assessment methods in Chapter 13). Two broad categories of restricted-response test item formats can be used to measure instructional goals and objectives:

- *Objective test items,* which include true-false, fill-in-the-blank, multiple-choice, and matching items.
- *Restricted essay items,* which pose a specific problem for which the student recalls information, organizes it in a suitable manner, derives a defensible conclusion, and expresses it within specified guidelines.

The following sections discuss some considerations that can increase the content validity of your restricted response tests.

Objective Test Items

Objective test items have four common formats: true-false, matching, multiple-choice, and completion (fill-in). How you write your objectives may predetermine the format, but in many instances you will have a choice between several item formats. In the following sections, we consider the construction and use of true-false, matching, multiple-choice, and completion items.

True-False Items

True-false items are popular with teachers because they are quick and easy to write, or at least they seem to be. True-false items do take less time to write than good objective items of any other format, but *good* true-false items are not so easy to prepare.

As you know from your own experience, every true-false item, regardless of how well or poorly written, gives the student a 50 percent chance of guessing the right answer correctly, even without his or her reading the item! In other words, on a 50-item true-false test, we would expect individuals who were totally unfamiliar with the content being tested to answer about 25 items correctly. Fortunately, ways exist to reduce the effects of guessing. Here are some:

- Encourage *all* students to guess when they do not know the correct answer. Because it is virtually impossible to prevent certain students from guessing, encouraging all students to guess equalizes the effects of guessing. The test scores will then reflect a more or less equal "guessing factor" *plus* the actual level of each student's knowledge. This will prevent test-wise students from having an unfair advantage over students who are not test-wise.
- Require revision of statements that are false. In this approach, you provide space at the end of the item for students to alter false items to make them true.

Usually the student is asked to first underline or circle the false part of the item and then add the correct wording, as in these examples:

T F High IQ students always get good grades.

often; tend to

T F Panama is north of Cuba.

south

T F September has an extra day during leap year.

February

With this strategy, you would award full credit only if the student's revision is correct. The disadvantage of such an approach is that more test time is required for the same number of items, and scoring time is increased.

The accompanying box, *Writing True-False Questions*, gives specific suggestions for writing these test items.

Matching Items

Like true-false, matching items are a popular and convenient testing format. Like good true-false items, however, good matching items are not easy to write. Imagine you are back in your ninth-grade American history class and the following item shows up on your test:

Directions: Match A and B

A	B	
1. Lincoln	a.	President during the twentieth century
2. Nixon	b.	Invented the telephone
3. Whitney	c.	Delivered the Emancipation Proclamation
4. Ford	d.	Only president to resign from office

5. Bell	e.	Black civil rights leader
6. King	f.	Invented the cotton gin
7. Washington	g.	Our first president
8. Roosevelt	h.	Only president elected for more than two terms

See any problems? Compare the problems you identified with the descriptions of faults that follow.

Homogeneity. The lists are not homogeneous. Column A contains names of presidents, inventors, and a civil rights leader. Unless these are specifically taught as a set of related people or ideas, this is too wide a variety for a matching exercise.

Order of Lists. The lists are reversed: column A should be in place of column B, and column B should be in place of column A. As the exercise is now written, the student reads a name and then has to read through all or many of the more lengthy descriptions to find the answer—a time-consuming process. It also is a good idea to introduce some sort of order—chronological, numerical, or alphabetical—to your list of options. This saves the student time.

Easy Guessing. Notice that there are equal numbers of options and descriptions. This increases the chances of guessing correctly through elimination. If there are at least three more options than descriptions, the chance of guessing correctly is reduced to one in four.

Poor Directions. The instructions are much too brief. Matching directions should specify the basis for matching. For example, “Column A contains brief descriptions of historical events. Column B contains the names of U.S. presidents. Indicate who was president when the historical event took place by placing the appropriate letter to the left of the number in column A.”

Multiple Correct Responses. The description “President during the twentieth century” has three correct answers: Nixon, Ford, and Roosevelt. Also, always include first and last names to enhance recall and avoid ambiguities. Here is a corrected version of the matching items we critiqued above:

Column A	*Column B*
_____ 1. Only president not elected to office.	a. Gerald Ford
_____ 2. Delivered the Emancipation Proclamation.	b. Thomas Jefferson
_____ 3. Only president to resign from office.	c. Abraham Lincoln
_____ 4. Only president elected for more than two terms.	d. Richard Nixon
	e. Franklin Roosevelt
	f. Theodore Roosevelt
	g. George Washington
	h. Woodrow Wilson

Notice that we now have complete directions, more options than descriptions, homogeneous lists (all items in Column A are descriptions of U.S. presidents and all items in Column B are names of presidents), and unambiguous alternatives.

The accompanying box, *Writing Matching Items*, contains some additional suggestions for writing these types of questions.

Multiple-Choice Items

Another popular item format is the multiple-choice question. Multiple-choice tests are more common in high school and college than in elementary school. Multiple-choice items are unique among objective test items because if properly written they enable you to measure some limited types of higher-level cognitive objectives.

However, multiple-choice items are more difficult to write than are the other types we have discussed so far. In the following sections, we discuss some common problems with multiple-choice items and provide specific suggestions for you to use when writing them. Following these suggestions will allow you to avoid inadvertently providing students with clues to the correct answer.

Stem Clue. The statement portion of a multiple-choice item is called the **stem,** and the answer choices are called *options* or **response alternatives.** A *stem clue* occurs when an identical or similar term appears in both the stem and an option, thereby clueing the test-taker to the correct answer. For example:

The free-floating structures within the cell that synthesize protein are called _____.

A. chromosomes
B. lysosomes
C. mitochondria
D. free ribosomes

In this item the word *free* in the option is identical to *free* in the stem. Thus, the wise test-taker has a good chance of answering the item correctly without mastery of the content being measured.

Grammatical Clue. Consider this item:

U. S. Grant was an _____.

A. cavalry commander
B. navy admiral
C. army general

D. senator

Most students would pick up on the easy grammatical clue in the stem. The article *an* eliminates options A, B, and D, because "*an* navy admiral," "*an* cavalry commander," or "*an* senator" are grammatically incorrect. Option C is the only one that forms a grammatically correct sentence. A way to eliminate the grammatical clue is to replace *an* with *a/an.* Similar examples are *is/are, was/were, his/her,* and so on. Alternatively, place the article (or verb, or pronoun) in the options list:

Christopher Columbus came to America in _____.

A. a car
B. a boat
C. an airplane
D. a balloon

Redundant Words/Unequal Length. Two very common faults in multiple-choice construction are illustrated in this item:

When 53 Americans were held hostage in Iran, _____.

A. the United States did nothing to free them
B. the United States declared war on Iran
C. the United States first attempted to free them by diplomatic means and later attempted a rescue
D. the United States expelled all Iranian students

The phrase "the United States" is included in each option. To save space and time, add it to the stem: "When 53 Americans were held hostage in Iran, the United States _____." Second, the length of options could be a giveaway: the correct option, C, is much longer than any of the others. Multiple-choice item writers have a tendency to include more information in the correct option than in the incorrect options. Test-

wise students know that the longer option is the correct one more often than not. Avoid making correct answers look different from incorrect options.

All of the Above/None of the Above. In general, use “none of the above” sparingly. Some item writers use “none of the above” only when no clearly correct option is presented. However, students catch on to this practice and guess that “none of the above” is the correct answer without knowledge of the content being measured. Also, at times it may be justified to use multiple correct answers, such as “both a and c” or “both b and c.” Again, use such options sparingly, because inconsistencies can easily exist among alternatives that logically eliminate some from consideration. Avoid using “all of the above,” because test items should encourage discrimination, not discourage it.

Higher-Level Multiple-Choice Questions

A good multiple-choice item is the most time-consuming type of objective test item to write. Unfortunately, most multiple-choice items are written at the knowledge level in the taxonomy of educational objectives. As a new item writer, you will tend to write items at this level, but you should learn to write multiple-choice items that measure cognitive objectives beyond the knowledge level. Following are some suggestions to make your higher-level multiple-choice questions more effective. In the next chapter we will show you how to measure cognitive objectives at the analysis, synthesis, and decision-making levels with performance-based assessments.

Use Justification to Assess Reasons Behind an Answer. Follow up on multiple-choice items with open-ended questions that ask students to specify why they chose their answers. This allows students to demonstrate knowledge at the comprehension level. For example:

Directions: Choose the most appropriate answer and cite evidence for your selection in the space below.

The principal value of a balanced diet is that it _____.

A. increases your intelligence

B. cures disease

C. promotes mental health

D. promotes physical health

E. improves self-discipline

What evidence from the text did you use to choose your answer?

Use Pictorial, Graphic, or Tabular Stimuli. Presenting pictures, drawings, graphs, or maps that the student must use to choose the correct answer to a multiple-choice question can require the student to think at least at the application level and may involve even higher cognitive processes (see Table 12.1). Also, such stimuli often can generate several higher-level multiple-choice items, as the questions below and Figure 12.1 illustrate:

Which of the following cities (identified by their grid locations on the accompanying map) would be the best location for a steel mill?

A. Li (3A)

B. Um (3B)

C. Cot (3D)

D. Dube (4B)

Approximately how many miles is it from Dube to Rag?

A. 100 miles

B. 150 miles

C. 200 miles

D. 250 miles

In what direction would someone have to travel to get from Wog to Um?

A. northwest

B. northeast

C. southwest

D. southeast

Use Analogies to Show Relationships Between Terms. To answer analogies correctly, students must not only know what the terms mean (knowledge) but they also must understand how they relate to each other (comprehension or application). For example:

Physician is to humans as veterinarian is to:

A. fruits

B. animals

C. minerals

D. vegetables

Require Application of Principles or Procedures. To test whether students comprehend the implications of a procedure, have them use the principle or procedure with new information or in a novel way. This requires them to do more than just follow the steps in solving a problem: They must also demonstrate an ability to apply their knowledge to a new context (application). Consider this example from a math test. The material covered was a division lesson on computation of ratios and proportions:

After filling his car's tank with 18 gallons of gasoline, Mr. Watts said to his son, "We've come 450 miles since the last fill-up. What gas mileage are we getting?" Which is the best answer?

A. 4 miles per gallon

B. 25 miles per gallon

C. Between 30 and 35 miles per gallon

D. It can't be determined from the information given

This item tests not only knowledge of division but also application skills.

The accompanying box, *Writing Multiple-Choice Items*, gives some specific guidelines to follow when you write items of this type.

Completion Items

Like true-false items, completion items are relatively easy to write. The first tests constructed by classroom teachers and taken by students often are completion tests. Like items of all other formats, there are good and bad completion items. Here are some suggestions for writing completion items:

- Require a single-word answer or a brief, definite statement. Avoid items so indefinite that they may be logically answered by several terms:

  Poor item: World War II ended in _____.

  Better item: World War II ended in the year _____.

- Be sure the item poses a specific question. An incomplete statement is often clearer than a question because it provides more structure for an answer.

  Poor item: Who was the main character in the story "Lilies of the Field?" _____.

Better item: The main character in the story “Lilies of the Field” was called _____.

- Be sure the answer is factually correct. Precisely word the question in relation to the concept or fact being tested. Make sure that the answer is included in the students’ text, workbook, or class notes.
- Omit only key words; don’t eliminate so many elements that the sense of the content is impaired.

  Poor item: The _____ of test item usually is graded more _____ than the _____ type.

  Better item: The multiple-choice type of test item is usually graded more objectively than the _____ type.
- Word the statement so the blank is near the end. This prevents awkwardly phrased sentences. For example:

  Poor item: In _____ , John F. Kennedy was elected president.

  Better item: John F. Kennedy was elected president in the year _____.
- If the question requires a numerical answer, indicate the units in which it is to be expressed (for example, pounds, ounces, minutes).

Advantages and Disadvantages of Objective-Item Formats

Table 12.3 summarizes the advantages and disadvantages of each of the objective item formats we have discussed: true-false items, matching items, multiple-choice items, and completion items.

Restricted Response Essay Items

In essay items, the student supplies, rather than selects, the correct answer. An essay test requires that the student compose a response to a question for which no *single*

response or pattern of responses can be cited as correct to the exclusion of all others. The accuracy and quality of a response to such a question often can be judged only by a person skilled in the subject area.

Like objective test items, essay items can be well constructed or poorly constructed. The well-constructed essay item tests complex cognitive skills by requiring the student to organize, integrate, and synthesize knowledge; to use information to solve novel problems; or to be original and innovative in problem solving. The poorly constructed essay item may require the student to do no more than recall information as it was presented in the textbook or lecture. Worse, the poorly constructed essay item may leave the learner unclear about what is required for a satisfactory response.

An essay item that allows the student to determine the length and complexity of a response is called an **extended-response essay** item. This type of essay is most useful at the analysis, synthesis, and evaluation levels of cognitive complexity. Because of the length of this type of item and the time required to organize and express the response, the extended-response essay is sometimes better assigned as a term paper, literary script, or research report. The extended-response essay often is of value both in assessing communication ability and in assessing achievement. We will have more to say about extended-response essays in the next chapter, on performance-based assessment.

An essay that poses a specific problem for which the student must recall information, organize it in a suitable manner, derive a defensible conclusion, and express it according to specific criteria is called a **restricted-response essay** item. The statement of the problem specifies limitations on the response that guide the student in responding and provides evaluation criteria for scoring. Following is an example of a well-written restricted-response essay item. Note how it specifies

exactly what information is required, how it should be organized, and how the response will be graded.

> List the major similarities and differences between U.S. participation in the Korean War and World War II, being sure to consider political, military, economic, and social factors. Limit your answer to one page. Your score will depend on accuracy, organization, and conciseness.

Using Restricted Essay Questions

Specific situations that lend themselves to restricted-response essay questions are those that require high-level thought processes, those in which it is necessary to set precise time or length limits for responses, and those in which you wish to tap more than one learning objective. The following criteria will help you write valid restricted-response essay questions.

- The instructional objectives for essay questions should specify higher-level cognitive processes. In other words, your aim in requiring essay answers is for students to supply information, not just recall information. The processes of analysis and synthesis often cannot be measured with objective items.
- Restricted-response essay tests are appropriate in situations where relatively few areas of content are to be tested. If you have 30 students and design a test with six restricted-response essays, you will spend a great deal of time scoring. Use restricted essays when class size is small, or use them in conjunction with objective items. Design your test blueprint to include a number of objective questions and only one or two essays.
- Essay responses help to maintain test security. If you are afraid test items will be passed on to future students, it is best to use an essay test format. In

general, a good essay test takes less time to construct than a good objective test.

- Restricted-response essays are a good choice when you want to test any of the following learning objectives:

  Analyze relationships

  Compare positions

  State necessary assumptions

  Identify appropriate conclusions

  Explain cause-and-effect relations

  Formulate hypotheses

  Organize data to support a viewpoint

  Point out strengths and weaknesses

  Integrate data from several sources

  Evaluate the quality or worth of an item, product, or action

The accompanying box, *Writing Restricted-Response Essay Questions*, provides specific suggestions that will help you write good essay questions.

Scoring Essays

Essays are difficult to score consistently. That is, the same essay answer may be given an A by one scorer and a B or a C by another scorer. Or the same answer may be graded A on one occasion, but B or C on another occasion by the *same* scorer (Coffman, 1972). What can you do to avoid such scoring problems?

Write Good Essay Items. Poorly written questions are one source of scorer inconsistency. Questions that do not specify response length are another. In general, longer essay responses are more difficult to score consistently than shorter responses. This is due to student fatigue and consequent mechanical errors as well as to a

tendency for grading criteria to vary from response to response or, for that matter, from page to page, or even paragraph to paragraph within the same response.

Use Several Restricted-Response Items. Rather than a single lengthy restricted-response essay, use several smaller essays. This will provide students a greater opportunity to show off their skills and a greater variety of criteria to respond to.

Use a Predetermined Scoring Scheme. All too often, essays are graded without the scorer having specified in advance what he or she is looking for in a "good" answer. If you do not specify the criteria beforehand, your scoring consistency will be greatly reduced. If these criteria are not readily available (in written form) for scoring each question, the criteria themselves may change (you may grade harder or easier after scoring several papers, even if the answers are similar). Or your ability to keep these criteria in mind will be influenced by fatigue, distractions, frame of mind, and so on. Because we all are human, we all are subject to these factors.

We will go into greater detail about the procedures involved in scoring open-ended responses in the next chapter, on performance assessment. Now we turn to potential questions of the content validity and reliability of restricted-response essay questions, and consider some ideas about how to solve these problems.

Some Unresolved Problems

Content Validity. Although multiple-choice questions and restricted-response essay questions can be written in a manner that requires thinking and problem solving, the learner's response may not always indicate that either of these processes took place. Depending on the degree of similarity between what was taught and what was tested, and on how clearly the teacher demonstrated the thinking and problem-solving skills, the learner could have arrived at the answer in a variety of ways—some

involving thought and inquiry, some simply involving recall. For example, suppose that you posed the question “What are the reasons trade agreements do or do not work?” just after covering the details of the North American Free Trade Agreement (NAFTA). This might evoke a genuine problem-solving response involving some analysis, or simply a paraphrase of the conditions required by NAFTA given in class. Therefore, even though you follow a test blueprint and write questions to match the performance desired, the actual thought process students use may differ from what you intended.

The most valid way to measure higher thought processes involving inquiry, problem solving, and decision making is to observe the learner’s performance related to the skill or behavior in question. In other words, assessment of higher thought processes seldom can be separated from the application of those processes to real-world problems. We will address this issue and provide examples in the following chapter.

Reliability Versus Validity. Teachers must be concerned about a test’s reliability as well as its content validity. Recall that *validity* is the degree to which a test measures the traits, abilities, or skills for which it was intended. A test’s *reliability,* on the other hand, is the degree to which the test dependably or consistently measures that trait, ability, or skill.

One way to think about reliability is in terms of an assessment instrument with which we are already familiar: the bathroom scale. A bathroom scale is *valid* if it measures how many pounds you weigh rather than how tall you are. It is *reliable* if it registers your 150-pound weight every time you step on it. If you weigh 150 pound but the scale sometimes reads 145, other times 160, and still other times 155, you have an unreliable scale. A reliable test, therefore, is one that dependably and

consistently gives approximately the same score regardless of how many times you take it (assuming you don't improve with practice).

Reliability is a quality we want for all our tests, but it is sometimes hard to attain. For example, unclear test questions may be interpreted differently by the learner each time they are read. In such cases, the learner's score will not accurately reflect what she knows or can do. Unclear instructions may produce test scores that are similarly unreliable. Poorly constructed tests often produce scores that are unreliable because both the instructions and the items are poorly worded. Also, keep in mind that when a scoring guide is not used for essays, the essays may be scored subjectively, further decreasing the test's reliability.

Several steps can be followed to increase test reliability. They are summarized in the accompanying box, *Improving the Reliability of Your Tests.*

Reporting Learner Progress:
Your Grading System

Consider the following comments about grading made by these first-year teachers:

> Of all the paperwork, grading is the nitty-gritty of teaching for me. Students take their grade as the bottom line of your class. It is the end-all and be-all of the class. To me, a grade for a class is, or at least should be, a combination of ability, attitude, and effort. Put bluntly: How do you nail a kid who really tried with an F? Or how do you reward a lazy, snotty punk with an A? (Ryan, 1992, p. 4)

> I have been amazed at comments they make about grades. Those with A's ask if they are flunking; others who rarely hand in assignments ask if I think they'll get a B. They make no connection between their own efforts and the grade they receive....Of course, there is some truth in their view of the

> arbitrariness of grading. But I don't like the powerlessness it implies. "I don't give you your grade," I tell them. "You do!" (Ryan, 1992, p. 96)
>
> Grading is still kind of a problem with me....I try not to play favorites [even though I have them]—I don't like S's personality...I do like J's, isn't it unfair? You want to be easier on someone you like. Or harder on someone you don't....I wouldn't mind taking a class on grading. I think grading could be hit much harder in college....I just don't know what to do, kind of. (Bullough, 1989, p. 66)

These comments from first-year teachers highlight the confusion, uncertainty, and even fear that surround the responsibility of assigning grades. Most of these negative feelings are related to two facts about grades: (1) They become part of a learner's permanent record viewed by parents, teachers, and future employers; and (2) while there are many procedures for assigning grades, there is little research to support one over another. Consequently, teachers' choices of grading procedures reflect largely their own values, past experiences, and the norms and traditions of the schools in which they work. Their awareness of the significance of grades and the sometimes arbitrary decisions on which they are based understandably gives new teachers cause for concern.

In this section we will identify the most important of the decisions you must make when assigning grades and provide a number of specific recommendations. Although your choice of a grading procedure will reflect your own values, it should also reflect the following beliefs:

- The primary purpose of assigning a grade should be to communicate what the student has learned from your instruction.

- Grades should be based on a variety of indicators of learning, including both written and oral formats, process criteria as well as product criteria, and formal and informal assessments.
- The criteria you use to make up your grading procedure should always be known to learners and parents.

The Purpose of a Grade

Most educators agree that the purpose of assigning a grade in reading, social studies, math, science, or any other academic subject should be to communicate how well the learner achieved your instructional goals and objectives (Hills, 1981; Kubiszyn & Borich, 1996; Thorndike, Cunningham, Thorndike, & Hagen, 1991; Tuckman, 1988). This will be accomplished if two criteria are met: (1) Your teaching focuses on your goals and objectives and (2) your assessments measure what you teach. Therefore, the most important consideration in deciding what grades to assign is how well learners have achieved your goals and objectives.

When the basis for a grade is linked to a standard of mastery or achievement you have determined, the grade is said to be criterion-referenced (Thorndike et al., 1991). An example of **criterion-referenced grading** would be to assign a grade of A to learners who averaged 90 percent across all your tests, a grade of B to those who averaged 80 to 89 percent, a grade of C to those who averaged 70 to 79 percent, and so on. In criterion-referenced grading, the grade is compared with a fixed standard of achievement.

If, however, the basis for assigning a grade is the comparison of a learner's performance with the performance of other learners, a norm-referenced grading system is being used. **Norm-referenced grading** refers to a procedure for assigning grades or scores based on how one learner's achievement compares to the achievement of other learners. In one such system, called "grading on a curve," the

teacher decides, for example, that the top 10 percent of learners will get As, the next 20 percent Bs, the next 40 percent Cs, the next 20 percent Ds, and the bottom 10 percent Fs.

"Normal curve grading" refers to a system whereby the percentages of As, Bs, Cs, Ds, and Fs are set in reference to the bell-shaped distribution of scores called the *normal curve,* which we studied in Chapter 11. One version of this system, illustrated in Figure 12.2, uses the standard deviation (also studied in Chapter 11) to indicate the particular percentage cutoff chosen for each letter grade, assuming that the test scores are normally distributed as shown.

Figure 12.3, in contrast, shows the kind of distribution of scores that typically occurs in criterion-referenced grading. In criterion-referenced grading, divisions between As, Bs, Cs, Ds, and Fs are made on the basis of the number or percentage of test items answered correctly as established by the teacher—not by how well others perform on the test, as in the case of a norm-referenced test. For example, in Figure 12.3, this teacher decided that getting 90 percent or more of the items correct on the test deserved the grade of A, getting 80 to 89 percent of the items correct deserved the grade of B, and so on. If the instruction has been effective, more students will score at the high end of the distribution with the criterion-referenced approach than with the norm-referenced approach. This creates a *negatively skewed* distribution, with the longer tail of the curve extending to the left, as shown in Figure 12.3.

An assigned grade in a criterion-referenced grading system is more directly influenced by the degree to which a student learned what was taught than by how many others received the same or a higher grade. In other words, the standard for assigning grades in criterion-referenced grading is *absolute*—established by the teacher—as opposed to *relative*—as established by how other students performed.

On What Should a Grade Be Based?

There are many ways for learners to demonstrate achievement consistent with your goals and objectives. However, teachers, particularly at the secondary level, tend to rely predominantly on paper-and-pencil tests and written assignments to measure learner achievement. This reliance on the written over the spoken word may place certain learners at a disadvantage, particularly in culturally diverse classrooms (Bennett, 1990). By using a variety of assessment techniques, and balancing your assessments between paper-and-pencil objective tests and performance assessments, you are more likely to be fair to all learners. We discuss the subject of performance assessments in depth in Chapter 13 and that of culture-fair instruction in Chapter 15.

Including a variety of indicators of achievement raises the question of the importance or weight each component carries in the overall final course grade. **Grade weighting** involves assigning degrees of importance to the different performance indicators that are combined to determine a grade. For example, a teacher may decide that homework and classwork count for 20 percent of the grade, quizzes 20 percent, performance assessments 30 percent, formal tests 20 percent, and a journal or notebook 10 percent.

The more weight a grading component plays in your final grade, the more accountable for that component learners will be (Emmer, Evertson, Clements, & Worsham, 1994). Learners are more likely to make a special effort to do classwork and homework well when they represent a significant portion of the final class grade. Additionally, you will want to give more weight to the indicators that you judge to be the most valid and reliable indicators of achievement. Formal tests, for example, usually are better indicators that students have mastered basic facts and rules than either responses to oral questions or class participation.

Finally, be aware that if the different indicators of achievement are graded on different point scales (for example, tests are worth 100 points, quizzes 10 points,

projects 50 points, homework 20 points, and so on), simply weighting them and then combining them into a final score may not give the components the importance they deserve. The accompanying box, *Using Grading Formulas,* illustrates three formulas commonly used in schools that attempt to balance these concerns.

Making Public Your Decisions About Grading

Learners are often uncertain about their grades at any given point in the school year and confused about how the grade is determined. Likewise, parents are often unaware of what goes into a grade and the relative importance placed on homework, participation, formal tests, and other components.

Once you have developed your grading system, you should present it to your learners, both orally and in a handout that can be sent home. Use numerical examples in both cases. While learners in the early elementary grades may not understand all of the complexities of assigning a grade, their parents will. At both the elementary and the secondary levels, a handout of your grading procedures will communicate a feeling of your accountability to parents and your values regarding achievement and grades.

Summing Up

This chapter introduced you to objective and essay tests. Its main points were these:

- Content validity is the degree to which a test covers all the content that was taught with the degree of emphasis in which it was taught.
- Two general types of test item formats are restricted-response and flexible-response. Restricted-response item formats include true-false, fill-in, multiple-choice, matching, and restricted essays. Flexible-response item formats include extended essays, term papers, research reports, and other performance-based assessments.

- Restricted-response test formats are best suited for measuring behavior at the knowledge, comprehension, and application levels; flexible-response test formats are best suited for measuring behavior at the analysis, synthesis, and decision-making levels.
- A test blueprint is a table that identifies the behaviors and content to be tested. The number and types of test questions are provided for each behavior, by content area, to indicate the instructional emphasis given to each.
- True-false questions are quick and relatively easy to construct but tend to emphasize memorization and guessing.
- Matching questions are suitable for measuring associations between facts but tend to ask trivial information and, depending on the number of alternatives, may not be adaptable to commercial answer sheets.
- Multiple-choice questions are versatile in measuring objectives from the knowledge to application levels but can be time-consuming to write and, if not carefully written, can have more than one defensible answer.
- Completion items reduce guessing but can be difficult to score, leading to multiple defensible answers.
- Restricted-response essay tests can measure behavior at higher levels of behavioral complexity but can be time-consuming to grade and require a scoring key or model answer prepared in advance.
- Criterion-referenced grades are based on standards of mastery or achievement that the teacher has determined. Norm-referenced grades are based on the relative performance of others who have taken the test.

For Discussion and Practice

*1. Identify three problems with Mr. Steele's grading practices that became apparent during his conversation with Ms. Dalton. How could each be remedied?

2. Provide one example each of how a properly constructed test could

reactivate previously learned skills and knowledge.

influence students' choice of study patterns and learning strategies.

establish a realistic picture of students' own abilities and competence.

*3. Provide one example each of how you might increase the fairness of your tests by making allowances for (a) gender, (b) culture/ethnicity, and (c) ability.

*4. Explain in your own words why a test, even when well constructed, will not be a perfect assessment of your learners' performance and progress.

*5. What would be an indirect assessment—or behavioral sign—for measuring the following objectives:

How to get the correct change in a convenience store

How to determine the maximum price you could pay for gasoline to travel two thousand miles if you had only $100 and your car gets 32 miles to the gallon

How to determine the angle of a roof with a 3-foot slope for every 10 feet of surface area.

*6. In your own words, describe the concept of test validity. What would be the result on your learners' scores of using a test that is not valid?

*7. What three traps do teachers generally fall into when constructing paper-and-pencil tests? In your opinion, which do you believe occurs most frequently?

8. Prepare a test blueprint for a subject of your own choosing that includes at least five content areas and three levels of behavior. Indicate the number and type of test items you will include in each cell of your blueprint and the totals for each content area and behavior.

*9. Provide a sequence of steps that can be used during the item-writing process to ensure a match between your objectives and your test items.

*10. What does a restricted essay test question require of the learner? Provide an example in your teaching area.

11. Write one multiple-choice test question that asks for justifications and one true-false question that asks for revisions. Provide written instructions to your students for completing each type of question.

*12. If you taught three classes of the same subject, each with 30 students to whom you have given two essay questions requiring approximately 150 words each to answer, how many words would you be required to read? About how long do you think it would take you? Assume you will read the essays at 300 words per minute.

13. Prepare a higher-level multiple-choice question in your teaching area requiring the application of principles or procedures.

*14. Describe in your own words the concepts of validity and reliability. What would be the effects on your learners' scores of an unreliable test?

*15. What six steps can help you increase the reliability of your tests? In your opinion, which is the most difficult to achieve?

*16. Contrast the way in which a grade would be determined using a norm-referenced test with the way it would be determined using a criterion-referenced test.

17. Choose one formula for computing and weighting grades from the box titled *Using Grading Formulas* (p. 418) and show with specific assessments how you would apply it during a six-week grading period in your classroom.

Suggested Readings

Kubiszyn, T., & Borich, G. (1996). *Educational testing and measurement* (5th ed.). New York: HarperCollins. Classroom teachers will find particularly helpful the discussions of essay grading, performance assessments, and the measurement of learner attitudes.

Popham, W. (1990). *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall. This text provides comprehensive treatment of issues surrounding testing and grading. It offers numerous practical examples of grading systems as well as insightful critiques of traditional grading procedures.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York: Macmillan. This text devotes extensive coverage to planning, classroom tests, and rules for writing a variety of test items. It also gives a thorough treatment to standardized tests.

Test fairness. A pattern of evaluation in which the teacher provides an authentic assessment of what has been taught, motivates learners toward higher levels of

effort, is sensitive to learner differences, accurately communicates performance and progress to learners and other parties, and efficiently uses teacher and learner time and effort.

How can I make sure that my assessments are fair to students?

Tests that measure what students learn and practice in the classroom can produce high levels of effort and student success.

Test validity. The capacity of a test to measure what it says it is measuring.

Content validity. A measure of the degree to which a test covers all the content that was taught in the manner in which it was taught.

Restricted-response tests. Assessment methods that limit the range of possible answers, such as true-false or multiple-choice tests, and are usually intended to test knowledge, comprehension, and application behaviors.

Flexible-response tests. Tests that measure higher thought processes such as analysis, synthesis, and decision-making behaviors usually through performance-based assessments.

Table 12.1

Classification of Behaviors in the Cognitive Domain by Test Type

Type of Test	Level of Behavioral Complexity	Expected Student Behavior	Instructional Process	Key Words
Restricted-response	Knowledge (remembering)	Student is able to remember or recall information and recognize facts, terminology, and rules	Repetition Memorization	define describe identify
	Comprehension (understanding)	Student is able to change the form of a communication by translating and rephrasing what has been read or spoken	Explanation Illustration	summarize paraphrase rephrase
	Application (transferring)	Student is able to apply the information learned to a context different from the one in which it was learned	Practice Transfer	apply use employ
Flexible-response	Analysis (relating)	Student is able to break down a problem into its component parts and draw relationships between the parts	Induction Deduction	relate distinguish differentiate
	Synthesis (creating)	Student is able to combine parts to form a unique or novel solution to a problem	Divergence Generalization	formulate compose produce
	Evaluation (decision	Student is able to make decisions about the value or worth	Discrimination Inference	appraise decide

making)	of methods, ideas, people, or products according to expressed criteria	justify

Paper-and-pencil tests can be valid and reliable measures of what students know. They need to be designed with instructional goals and objectives in mind.

Test blueprint. A table used to identify the type of behavior and content to be tested.

How can I make sure my classroom tests measure what I teach?

Table 12.2

Test Blueprint for a Physical Science Test

Topics	Memory for Facts and Terms 25%	Understanding of Concepts 40%	Application and Analysis of Data 35%	Totals
Weathering (15%)	5 fill-ins (1 pt each)	5 multiple-choice (2 pts each)		15 pts
Physical weathering (15%)	5 true-false (1 pt each)	5 multiple-choice (2 pts each)		15 pts
Volcanic activity (25%)	5 fill-ins (1 pt each)	5 multiple-choice (2 pts each)	1 essay (10 pts)	25 pts
Folded structures (15%)	5 fill-ins (1 pt each)		1 essay (10 pts)	15 pts
Faults (30%)	5 fill-ins (1 pt each)	5 multiple-choice (2 pts each)	1 essay (15 pts)	30 pts
TOTALS	25 pts	40 pts	35 pts	100 pts

How can I write test questions that require learners to use the same thought processes I emphasized during my instruction?

Applying Your Knowledge:

Matching Test Questions to
Instructional Objectives

- Have the goals and objectives of each of your lessons in front of you as you write or select test questions. Do not rely on memory to tell you if you intended your learners to memorize or to apply the processes involved in physical weathering.
- Take the learner's perspective. Reread each test question and ask yourself, "What thought process or intellectual skill is needed to answer this question correctly?" That is, at what level of behavioral complexity is the learner being asked to respond? Have you asked students to list the steps in a process, to compare that process to another, or to analyze its effects?
- Finally, refer back to the objective that corresponds with this question and ask, "Does this question reflect the level of behavioral complexity specified in the objective?" "Does the question reflect what I actually taught in the classroom?"

How do I choose among test items formats for an objective test?

How do I write objective test questions that accurately measure what I have taught?

Applying Your Knowledge:

Writing True-False Questions

- Tell students clearly how to mark *true* or *false* (for example, circle or underline the *T* or *F*) before they begin the test. Write this instruction at the top of the printed test as well.

- Construct statements that are definitely true or definitely false, without qualifications. If the item is true or false on the basis of someone's opinion, make sure the source is part of the item. For example, "According to the head of the AFL-CIO, workers' compensation is below desired standards."
- Write true and false statements that are approximately the same length—avoid the mistake of adding so many qualifying phrases that true statements are almost always longer than false ones. Similarly, include approximately equal numbers of true and false items.
- Avoid using double-negative statements. They take extra time to decipher and are difficult to interpret. For example, avoid statements such as "It is not true that addition cannot precede subtraction in algebraic operations." Instead write "Addition can precede subtraction in algebraic operations."
- Avoid terms denoting indefinite degree (for example, *large, long time, regularly*), or absolutes (*never, only, always*). These often cue students that a statement must be false. For example, "Congress and the President always cooperate to produce the federal budget."
- Avoid placing items in a systematic pattern that some students might detect (for example, true-true-false-false, true-false-true-false, and so on).
- If you use statements directly from the textbook, make sure that you are not taking them out of context.

Applying Your Knowledge:

Writing Matching Items

- Keep both the descriptions list and the options list short and homogeneous. They should fit together on the same page. Title the lists to ensure homogeneity (e.g.,

Column A, Column B) and arrange the options in a logical (e.g., alphabetical) order.

- Make sure that all the options are plausible distractors (wrong answer choices) for each description to ensure homogeneity of lists. In other words, make them logically parallel: don't include one famous general in a list of U.S. presidents, or one U.S. president in a list of foreign heads of state.
- The descriptions list should contain the longer phrases or statements, while the options should consist of short phrases, words, or symbols.
- Number each description (1, 2, 3) and letter each option (a, b, c,).
- To reduce the effects of guessing, include more options than descriptions.
- In the directions, specify the basis for matching (for example, "Match the civil rights leaders' names with the demonstrations they led," or "Match the procedure with the step of the scientific process in which you would use it"). Also, be sure to specify whether students can select each option more than once.

Stem. The statement portion of a multiple-choice question.

Response alternatives. The answer-choices portion of a multiple-choice question.

How do I write multiple-choice items that measure higher-order thinking skills?

Figure 12.1
Use of a pictorial stimulus to measure a higher-level cognitive process.

Applying Your Knowledge:

Writing Multiple-Choice Items

- Be sure that there is one—and only one—correct or clearly best answer.
- Be sure that all wrong answer choices (*distractors*) are plausible. Eliminate unintentional grammatical clues, and keep the length and form of all the answer choices equal. Rotate the position of the correct answer from item to item randomly.
- Use negative questions or statements only if the knowledge being tested requires it. In most cases it is more important for the student to know what the correct answer *is* rather than what it is *not.*
- Include three to five options (two to four distractors plus one correct answer) to optimize testing for knowledge, comprehension, or application rather than encouraging guessing.
- Use the option “none of the above” sparingly and only when all the answers can be classified unequivocally as wrong.
- Avoid using “all of the above,” especially as the correct answer. It makes it easy for students who have only partial information to guess the correct answer.

Extended-response essay. An essay question that allows the student to determine the length and complexity of a response; it is a good means of assessing communication ability as well as achievement.

Table 12.3

Advantages and Disadvantages of Various Objective Item Formats

True-False Items

Advantages	**Disadvantages**
Tend to be short, so more material can be covered than with any other format; thus, use T-F items when extensive content has been covered.	Tend to emphasize rote memorization of knowledge (although complex questions sometimes can be asked using T-F items).

Faster to construct (but avoid creating an item by taking statements out of context or slightly modifying them).	They assume an unequivocally true or false answer (it is unfair to make students guess at your criteria for evaluating the truth of a statement).
Scoring is easier (tip: provide a "T" and an "F" for them to circle, because a student's handwritten "T" or "F" can be hard to decipher).	Allow and may even encourage a high degree of guessing (generally, longer examinations compensate for this).

Matching Items

Advantages	**Disadvantages**
Simple to construct and score.	Tend to ask trivial information.
Ideal for measuring associations between facts.	Emphasize memorization.
Can be more efficient than multiple-choice questions because they avoid repetition of options in measuring association.	Most commercial answer sheets can accommodate only five options, thus limiting the size of a matching item.
Reduce the effects of guessing.	

Multiple-Choice Items

Advantages	**Disadvantages**
Versatile in measuring objectives, from the knowledge level to the application level.	Time-consuming to write.
Since writing is minimal, considerable course material can be sampled quickly.	If not carefully written, can have more than one defensible correct answer.
Scoring is highly objective, requiring only a count of correct responses.	
Can be written so students must discriminate between options varying in correctness, avoiding the absolute judgments of T-F tests.	
Reduce effects of guessing.	
Amenable to statistical analysis, so you can determine which items are ambiguous or too difficult (see Kubiszyn & Borich, 1996, Chapter 8).	

Completion Items

Advantages	**Disadvantages**
Question construction is relatively easy.	Encourage a low level of response complexity.
Guessing is reduced because the question requires a specific response. multiple	Can be difficult to score (the stem must be general enough not to communicate the answer, leading unintentionally to defensible answers).

Less time is needed to complete than multiple-choice items, so more content can be covered.	Very short answers tend to measure recall of specific facts, names, places, and events instead of more complex behaviors.

Restricted-response essay. An essay that poses a specific problem for which the student must recall proper information, organize it in a suitable manner, derive a defensible conclusion, and express it according to specific criteria.

How do I write essay questions that accurately measure what I have taught?

Well-written essay questions give learners clearly defined tasks and explicitly described standards for answering correctly. Answers to questions written in this manner are easier to rate or score.

Can essay tests be scored reliably?

Applying Your Knowledge:

Writing Restricted-Response Essay Questions

- Be clear about what mental processes you want the student to use before starting to write the question. Refer to the mental processes required at the various levels in the taxonomy of educational objectives (see Table 12.1). For example, if you want students to apply what they have learned, determine what mental processes would be needed in the application process.

  Poor item: Criticize the following speech by our president.

  Better item: Consider the following presidential speech. Focus on the section dealing with economic policy and discriminate between factual statements and opinions. List these statements separately, label them, and indicate whether each statement is or is not consistent with the President's overall economic policy.

- Make sure that the question clearly and unambiguously defines the task for the student. Tasks should be explained either in the overall instructions at the beginning of the test or in the test items themselves. Clearly state the writing style required (for example, scientific versus descriptive prose), whether spelling and grammar will be counted, and whether organization of the response will be an important scoring element. Also, indicate the level of detail and supporting data required.

  Poor item: Discuss the value of behavioral objectives.

  Better item: Behavioral objectives have enjoyed increased popularity in education over the years. In your text and in class the advantages and disadvantages of behavioral objectives have been discussed. Take a position for or against the use of behavioral objectives in education and support your position by using at least three of the arguments covered in class or in the text.

- Begin restricted-response essay questions with such words or phrases as *compare, contrast, give reasons for, give examples of, predict.* These words tell the student clearly what the answer must contain. Do not begin with such words as *what, who, when,* and *list,* because these words generally lead to tasks that require only recall of information.

  Poor item: List three reasons behind America's withdrawal from Vietnam.

  Better item: After almost 20 years of involvement, the United States withdrew from Vietnam in 1975. Predict what might have happened if America had *not* withdrawn at that time.

- A question about a controversial issue should ask for, and be evaluated in terms of, the presentation of evidence rather than the position taken. You should not demand that students accept your specific conclusion or solution. You can, however, require students to present and use evidence to support their own conclusions, whatever they may be.

Poor item: What laws should Congress pass to improve the medical care of all citizens in the United States?

Better item: Some feel that the cost of all medical care should be borne by the federal government. Do you agree or disagree? Support your position with at least three logical arguments.

- Establish reasonable time or length limits for each answer. This helps students complete the entire test and also indicates the level of detail you require. Indicate the time limit either in the statement of the problem or close to the question number.

Applying Your Knowledge:

Improving the Reliability of Your Tests

- Write test instructions and questions in simple, uncomplicated language. This way, learners won't have to struggle to untangle complicated phrases or clauses. When in doubt, have a student in a lower grade read the item and tell you what the question means.
- Include enough questions on the test to adequately cover all that the learner is expected to know. The more items, the greater the sample of performance obtained and the higher the test reliability.
- Allow students sufficient time to take the test. A rushed student is more likely to make foolish errors, which will lower the reliability of the test.
- Make sure the testing conditions (temperature of the room, noise level, seating arrangements) are conducive to maximum performance. If the testing situation is

uncomfortable, learners are more easily distracted and less likely to demonstrate what they really know.

- Follow a test blueprint so that you adequately sample all that has been taught—and all that the learner knows. The test blueprint will enable you to construct enough test items to adequately measure what you've taught.
- Write objective questions that have easily identifiable right and wrong answers. When using restricted-response essay questions, prepare a model answer or scoring guide before grading. This way the correct answer will be scored "right" every time, or a "good" answer given a high rating, because it reflects your scoring guide, not your mood or temperament at the time of grading.

Should I base my grades on how a learner's achievement compares with the achievement of other learners, or should I base them on a standard of mastery that I determine?

Criterion-referenced grading. The linking of grades to a standard of mastery or achievement.

Norm-referenced grading. The assignment of grades or scores based on how one learner's achievement compares with the achievement of other learners.

Figure 12.2

A normal curve, illustrating examples of percentages for distributing grades on a norm-referenced test.

Figure 12.3

A negatively skewed curve, illustrating the distribution of test scores and grades on a criterion-referenced test.

How do I combine different indicators of achievement, such as classwork, homework, quizzes, and projects, into a final grade?

Grade weighting. Assigning different degrees of importance to different performance indicators that are then combined into a grade.

Students are more likely to devote effort to homework assignments if they know that homework will make up a significant portion of their final grade.

Report cards are one of the principal means by which parents learn about their children's achievement. Learners and their parents should know on what the grade was based and how the grade was computed.

Applying Your Knowledge:

Using Grading Formulas

Following are examples of three commonly used grading formulas.

Example #1: The One, Two, Three Times Plan

One Time: All grades recorded for *homework* and *class work* are totaled and averaged. The average grade will count *one time* (one-sixth of the grade). For example, a student's homework and class work grades are:

84, 81, 88, 92, 96, 85, 78, 83, 91, 79, 89, 94=1040/12=86.6, or 87 average

Two Times: All of the *quizzes* are totaled and averaged. This average grade will count *two times* (one-third of the grade). For example:

82, 88, 80, 91, 78, 86=505/6=84.2, or 84 average

Three Times: All of the *tests* and *major projects* are totaled and averaged. This average will count *three times* (one-half of the grade). For example:

81, 91, 86=258/3=86 average

Final Grade: The final grade would be computed as follows:

87 (one time)+84+84 (two times)+86+86+86 (three times)=513/6=85.5=86

Example #2: The Percentages Plan

A teacher determines a percentage for each area to be included in the grade. For example, homework and class work will count for 20 percent; quizzes, 40 percent; tests and major projects, 40 percent. Using the same scores listed above, a student's grade would be computed as follows:

20 percent of the 86.6 for homework and class work=17.3

40 percent of the 84.2 for quizzes=33.7

40 percent of the 86 for tests and major projects=34.4.

To compute the final grade, add these three weighted averages:

17.3+33.7+33.4=85.4=85 as the final grade. (The average is different because the "weight" put on each area varies in the two examples.)

Example #3: The Language Arts Plan

A language arts teacher determines that four grades, those for publishing, goal meeting, keeping a journal, and daily process, each count for one-fourth (25 percent) of the grade.

The grade is computed as follows:

The publishing grade (issued at the end of the marking period)=88

The goal-meeting grade (issued at the end of the marking period)=86

The journal grades are 82+92+94+90+88+86=532÷6=88.7=89

The daily process grades are 78+82+86+94+94+91=525÷6=87.5=88

The six-weeks grade is 88+86+89+88=351÷4=87.75=88

Questions marked with an asterisk are answered in the appendix.