1

Achievement Testing and Instruction

Well-designed achievement tests support and reinforce other aspects of the instructional process. They aid both the teacher and the student in assessing learning readiness . . . monitoring learning progress . . . diagnosing learning difficulties . . . and evaluating learning outcomes. . . . The effectiveness of testing is enhanced by careful attention to the principles of test construction and to the use to be made of the results.

All types of instructional programs can benefit from the effective uses of achievement tests. In regular classroom instruction, achievement testing can serve a variety of useful functions. In individualized instruction, mastery learning programs, and the use of computer-assisted instruction, achievement testing is an indispensable procedure for monitoring and directing learning. Despite the widespread use of achievement testing and the important role it plays in instructional programs, however, many teachers receive little or no instruction in how to construct good achievement tests. This book is an attempt to correct that deficiency. It places major emphasis on those principles and procedures of test construction that are useful to classroom teachers. A guiding theme throughout the book is that achievement testing should be an integral part of the teaching-learning process.

An achievement test is *a systematic procedure for measuring a representative sample of learning tasks*. Although the emphasis is usually on measuring a set of *intended* learning outcomes, as defined by the instructional objectives, it should not be implied that testing be limited to the end of instruction. All too frequently, teachers view achievement testing as an end-of-unit or end-of-course activity that is done primarily for the purpose of assigning grades. Although this is one useful function of testing, it is just one of many. As with all teaching activities, the main

purpose of testing is to improve learning and it can contribute to this end in a number of ways.

TESTING IN THE INSTRUCTIONAL PROCESS

In order to realize the full potential of achievement tests as learning aids, it is necessary to make testing an integral part of the instructional process. Testing should be considered during the planning for instruction, and it should play a significant role in the various stages of instruction. From the beginning of instruction to the end there are numerous decisions that teachers must make. Testing can improve the effectiveness of many of these decisions by providing more objective information on which to base judgments.

Let us consider three types of decisions teachers need to make that can be aided by testing: (1) decisions at the beginning of instruction, (2) decisions during instruction, and (3) decisions at the end of instruction. Doing so will also help acquaint us with the names of the test types that are typically associated with each stage of instruction.

Beginning of Instruction (Placement Testing)

There are two major questions that teachers need to answer before proceeding with the instruction:

1. To what extent do the students possess the skills and abilities that are needed to begin instruction?
2. To what extent have the students already achieved the intended learning outcomes of the planned instruction?

Information concerning the first question can be obtained from *readiness* pretests. These are tests given at the beginning of a course or unit of instruction that cover those prerequisite skills necessary for success in the planned instruction. For example, a test of computational skill might be given at the beginning of an algebra course, or a test of English grammar might be given at the beginning of a German course. Students lacking in prerequisite skills could be given remedial work, or they could be placed in a special section that had lower prerequisites.

The second question can be answered by a *placement* pretest covering the intended learning outcomes of the planned instruction. This might very well be the same test that is given at the end of the instruction; preferably it should be another form of it. Here we are interested in determining whether students have already mastered some of the material we plan to include in our instruction. If they have, we might need to modify our teaching plans, encourage some students to skip particular units, and place other students at a more advanced level of instruction. The function of placement testing is summarized in Figure 1.1.

Placement testing is, of course, not always necessary. Teachers who have worked with students for some time may know their past achievements well enough

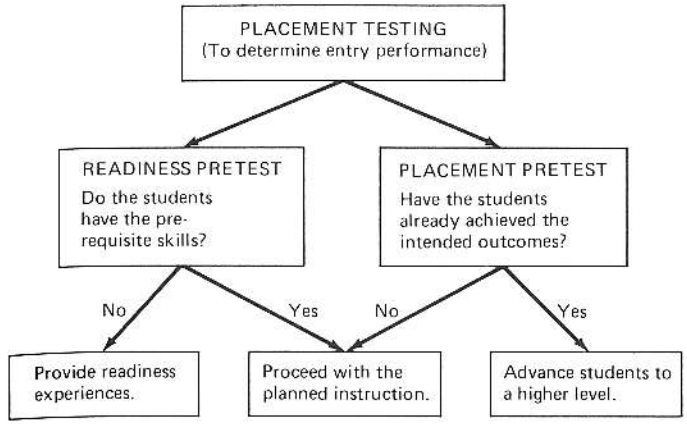


FIGURE 1.1. Simplified Model for the Instructional Role of Placement Testing.

that a pretest at the beginning of an instructional unit is not needed. In other cases a course or unit of instruction may not have a clearly defined set of prerequisite skills. Similarly, some areas of instruction may be so new to the students that it can be assumed that none of the students have achieved the intended outcomes of the planned instruction. Placement testing is probably most useful when the teacher is unfamiliar with the students' skills and abilities and when the intended outcomes of instruction can be clearly specified and organized in meaningful sequences. Under these conditions the placement test provides an invaluable aid for placing each student at the most beneficial position in the instructional sequence.

During Instruction (Formative and Diagnostic Testing)

During the instructional program our main concern is with the learning progress being made by students. Questions such as the following must be answered:

1. On which learning tasks are the students progressing satisfactorily? On which ones do they need help?
2. Which students are having such severe learning problems that they need remedial work?

Tests used to monitor student progress during instruction are called *formative* tests. Formative tests are typically designed to measure the extent to which students have mastered the learning outcomes of a rather limited segment of instruction, such as a unit or a textbook chapter. These tests are similar to the quizzes and unit tests that teachers have traditionally used, but they place greater emphasis on (1) measuring all of the intended outcomes of the unit of instruction, and (2) using the results to improve learning (rather than to assign grades). The purpose is to identify the students' learning successes and failures so that adjustments in instruction and learning can be made. When the majority of students fail a test item, or set of items,

the material is typically retaught in a group setting. When a minority of students experience learning failures, alternate methods of study are usually prescribed for each student (for example, reading assignments in a second book, programmed instruction, and visual aids). These corrective prescriptions are frequently keyed to each item, or to each set of items designed to measure a separate learning task, so that students can begin immediately after testing to correct their individual learning errors.

When a student's learning problems are so persistent that they cannot be resolved by the corrective prescriptions of formative testing, a more intensive study of the student's learning difficulties is called for. It is here that the *diagnostic* test is useful. This type of test typically includes a relatively large number of test items in each specific area with slight variations from one item to the next so that the cause of specific learning errors can be identified. The diagnostic test attempts to answer such questions as the following: Are the students having difficulty in addition because they don't know certain number combinations or because they don't know how to carry? Are the students' difficulties in reading German due to their inadequate knowledge of vocabulary or to their poor grasp of certain elements of grammar? Are the students unable to apply scientific principles to new situations because they don't understand the principles, because their knowledge of particular concepts is weak, or because the new situations are too unfamiliar to them? Thus, the diagnostic test focuses on the common sources of error encountered by students, so that the learning difficulties can be pinpointed and remedied.

Diagnosing learning problems is a matter of degree. The formative test determines whether a student has mastered the learning tasks being taught and, if not, prescribes how to remedy the learning failures. The diagnostic test is designed to probe deeper into the causes of learning deficiencies that are left unresolved by formative testing. Of course, this is not to imply that all learning problems can be overcome by formative and diagnostic testing. These are simply tools to aid in the identification and diagnosis of specific learning difficulties so that appropriate remedial steps can be taken. Diagnosing and remedying severe learning problems frequently requires a wide array of evaluative tools and the services of specially trained personnel. All we are attempting to do here is to show how formative and diagnostic tests can contribute to improved student learning during instruction. The model presented in Figure 1.2 summarizes the process.

End of Instruction (Summative Testing)

At the end of a course or unit of instruction we are concerned primarily with the extent to which the students have achieved the intended outcomes of the instruction. Questions such as the following must be answered:

1. Which students have mastered the learning tasks to such a degree that they should proceed to the next course or unit of instruction?
2. What grade should be assigned to each student?

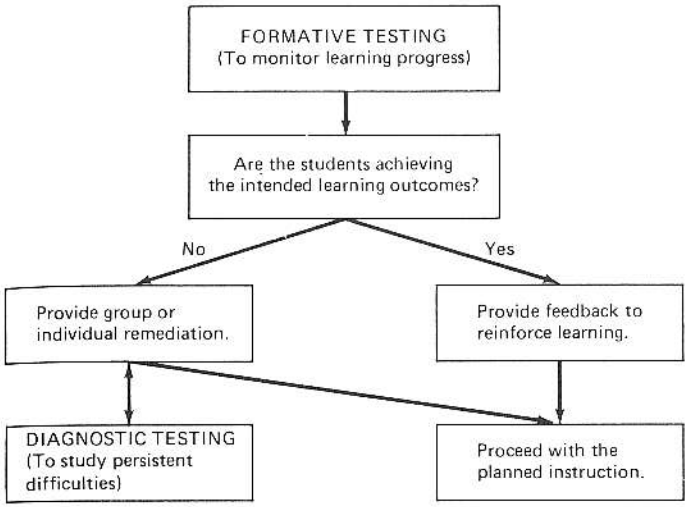
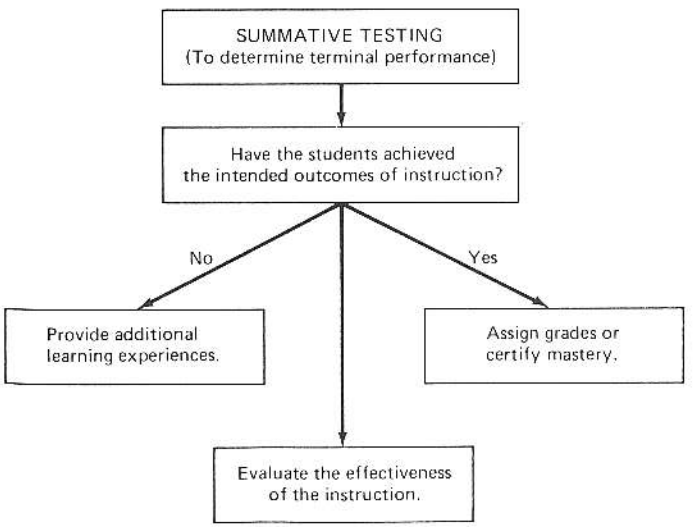


FIGURE 1.2. Simplified Model for the Instructional Role of Formative Testing.

The achievement test given at the end of a period of instruction for the purpose of certifying mastery or assigning grades is called a *summative* test. These tests are tyically broad in coverage and attempt to measure a representative sample of all of the learning tasks included in the instruction. Although the results are used primarily for grading, they can contribute to greater future learning by providing information for evaluating the effectiveness of the instruction. See Figure 1.3 for the summative testing model.

FIGURE 1.3. Simplified Model for the Instructional Role of Summative Testing.



OTHER WAYS TESTS CAN AID LEARNING

As noted in the last section, achievement tests can aid the teacher in making various instructional decisions having a direct influence on student learning. In addition, tests can aid student learning in a number of other ways.

Tests Can Aid Student Motivation

A carefully planned achievement testing program can have a direct influence on student learning by (1) providing them with short-term goals, (2) clarifying the types of tasks to be learned, and (3) providing feedback concerning their learning progress. Short-term goals are more motivating than admonitions that "some day you will find this information useful." An expected test stimulates learning activity and directs it toward the learning tasks to be covered by the test. Although this influence of testing is sometimes considered undesirable, as when the test measures only factual information, it need not be a negative influence. Its contribution to learning depends to a large extent on how faithfully our tests reflect all of the important outcomes of the instruction and how we use the results. For example, if the application of principles is stressed in our testing as well as in our teaching, we can expect students to direct greater efforts toward learning how to apply principles. Also, if the test results are reported to students as soon as possible, this feedback concerning their strengths and weaknesses in the application of principles will further clarify the nature of the task and indicate what changes are needed for effective performance. Thus, properly constructed tests can motivate students to work toward the instructional objectives of a course by arousing greater learning activity, by directing it toward the intended learning outcomes, and by providing prompt knowledge of results.

Tests Can Aid Retention and Transfer of Learning

Because tests tend to direct students' learning efforts toward the intended outcomes of instruction, they can be used as tools for increasing the retention and transfer of learning. In general, learning outcomes at the understanding, application, and interpretation levels are likely to be retained longer and to have greater transfer value than outcomes at the knowledge level. By including measures of these more complex learning outcomes in our tests, we can direct attention to their importance and provide reinforcing practice in the skills, applications, and interpretations we are attempting to develop. Thus, tests can be used to supplement and complement our teaching efforts in these areas and thereby increase the likelihood that the learning will be of greater permanent value to the students.

Tests Can Aid Student Self-Evaluation

All instruction should be directed toward helping individuals better understand themselves so that they can make more intelligent decisions. Periodic testing and feedback of the results can help students gain insight into what they can do well,

the misconceptions that need correction, and the degree of skill they have in various areas. Such information provides the students with a more objective basis for evaluating their strengths and weaknesses. Properly constructed tests tend to provide evidence of learning progress in such an objective and impartial way that the results can be accepted with little resistance or distortion. This assumes, of course, that the tests are properly constructed and are being used to improve learning rather than to threaten or label students. In the latter instance, self-evaluation is apt to be distorted by the psychological defense mechanisms an individual uses to maintain a positive self-image.

Tests Can Aid in Evaluating Instructional Effectiveness

Achievement test results can be used to evaluate the effectiveness of various aspects of the instructional process. For example, they can help determine the extent to which the instructional objectives were realistic, whether the methods and materials of instruction were appropriate, and how well the learning experiences were sequenced. When the majority of the students do poorly on the same test items, it may be the fault of the students but the difficulty is more likely to be found in the instruction. The teacher may be striving for learning outcomes that are unattainable by the students, may be using inappropriate materials, or may be using ineffective methods for bringing about the desired changes. An analysis of the students' responses to the test and the post-test discussion of the results should provide clues to the source of the instructional difficulty so that corrective steps can be taken.

MINIMUM COMPETENCY TESTING AND STUDENT LEARNING

One of the most wide-ranging attempts to use tests to improve student learning can be seen in the development of minimum competency testing programs. The majority of states has now mandated some type of state-wide minimum competency testing. These programs typically involve testing in the basic skills (reading, mathematics, and language usage), and in some cases in such areas as citizenship, consumer education, health education, and career planning. The tests frequently serve as a graduation requirement and students must demonstrate a satisfactory level of performance before receiving a high school diploma. Students who fail the tests are typically provided with remedial instruction and retesting. Repeated failure might result in the granting of a certificate of attendance instead of a diploma. In some states, competency tests are also given in the lower grades to monitor learning progress and identify weaknesses that can be remedied early in the students' educational development.

The minimum competency testing movement has focused national attention on the role of tests in improving learning. There is a real danger, however, that (1) the emphasis on *minimum* competency might direct so much effort toward low scoring students that it leads to the neglect of average and high achieving students, and (2) the narrow focus of the competency tests on basic and life skills might result

in less attention being given to other intended learning outcomes. These potential hazards can, of course, be partially offset by the use of classroom achievement tests that are carefully designed to measure all important outcomes of the instruction. How to construct such tests is the major focus of this book.

GUIDELINES FOR ACHIEVEMENT TESTING

Achievement tests are likely to make their greatest contribution to improved learning and instruction if test development is guided by a basic set of principles. The following principles provide a general framework for constructing achievement tests that are most effective in the teaching-learning process.

1. Achievement tests should measure clearly defined learning outcomes. The first step in test development is not the writing of test items but rather the clarification of what is to be measured. This includes considerations such as the following:

1. What are the intended learning outcomes to be measured? These are determined by the instructional objectives of the unit or course (e.g., knowledge of terms, knowledge of facts, understanding of concepts, application of principles).
2. What *specific* student responses will provide evidence that the learning outcomes have been achieved (e.g., "Students know terms" when they can (a) define the terms in their own words, (b) use the term in an original sentence, and (c) distinguish between terms that are similar in meaning)?

When the intended learning outcomes are clearly stated in terms of student performance it helps in selecting the types of test items to write. If we expect students to be able to "define terms in their own words," for example, we simply give them the terms and ask them to write definitions. A true-false or multiple-choice item would provide an inappropriate measure of this specific type of student performance.

2. Achievement tests should be concerned with all intended learning outcomes. In specifying the learning outcomes to be measured, attention should be given to all important outcomes. There is a temptation to focus on simple knowledge and skill outcomes because they are so easy to specify and measure. However, it we want to encourage students to go beyond the memorization of factual information, we must also include outcomes in the areas of understanding, application, and other complex types of learning. If we teach for understanding and application but do not test for it, students are not likely to give it much emphasis in their study. Letting students know that the tests will measure all levels of learning is one way that achievement testing can make a direct contribution to improved learning.

3. Achievement tests should measure a representative sample of instructionally relevant learning tasks. Testing is always a matter of sampling. We

seldom can ask all of the questions that we would like to ask in a test. Our instruction might cover hundreds of facts, terms, understandings, applications, and skills, but because of the time available for testing and other constraints we can include only enough items to measure a portion of them. It is the responsibility of the test maker to use a systematic procedure for obtaining a representative sample of test items relevant to the instruction. This involves proceeding from the intended outcomes of instruction (what students should be learning) to the domain of achievement for a particular test (what should be measured) and then to the test itself (what sample of test tasks to include). An instructionally relevant test requires that the instruction, the achievement domain, and the achievement test all be in close agreement (see Figure 1.4).

The selection of a representative sample of tasks for an achievement test can be easily illustrated with an example. Let's assume that an achievement domain specifies that students should be able to define 100 technical terms that have been included in the instruction. Because of time restrictions we can include only 20 items in our test. A fairly representative sample of these terms could be obtained, in this case, by simply placing the 100 words in alphabetical order and selecting every fifth word. Such careful sampling would enable us to generalize from the students' performance on the test sample to their probable performance on the larger domain of words that the test represents. Thus, if a student defined 90 percent of the words correctly on the 20-item test, we can assume that he or she probably could also define approximately 90 percent of the 100 words from which the sample was drawn. Since we are always interested in generalizing from the test sample to the larger achievement domain, the representativeness of the test sample is a key element in constructing an effective measure of achievement.

The procedures involved in obtaining a representative sample of tasks for testing purposes are typically much more involved than our simple example. Because achievement tests are commonly designed to measure a variety of learning tasks, more elaborate test specifications are needed. The procedure for preparing test specifications will be described in the following chapter.

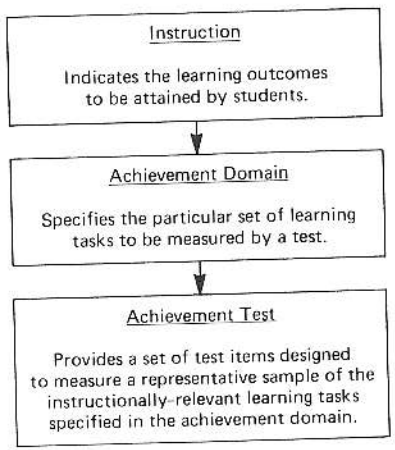


FIGURE 1.4.
Sequence in Constructing Instructionally Relevant Tests.

4. Achievement tests should include the types of test items that are most appropriate for measuring the intended learning outcomes. The learning outcomes for a course or unit of study specify the types of performance we are willing to accept as evidence of student achievement. The achievement test is simply a device for calling forth the specified performance so that judgments can be made concerning the extent to which learning has taken place. The key to effective achievement testing is to select the most appropriate item type and to construct it so carefully that it elicits the desired response.

In some cases a variety of item types may be needed to measure an intended learning outcome. By combining item types we can make up for a weakness in one by using the strength of the other. In determining students' ability to write, for example, logic would tell us that the best procedure is to have them write something. This obviously would provide the most direct measure of writing ability. However, in completing the assignment some students may select only those words they know well and can spell properly and write sentences that need only simple punctuation. Thus, a comprehensive measure of writing skills would require the use of supplementary measures of vocabulary, spelling, and punctuation. Using short-answer tests in these areas makes it possible to obtain systematic coverage of specific writing skills not possible with the writing project. Therefore, it is not a matter of either using a writing project or using short-answer tests. The two supplement and complement each other and the use of both results in more complete information.

5. Achievement tests should be based on plans for using the results. During the process of constructing achievement tests, consideration should be given to how the results will be used. For example, givng a pretest at the beginning of a unit or a course will be of little value unless plans are made to use the results to modify instruction as needed. Similarly, administering a diagnostic test presumes that remedial teaching will follow. In the use of other types of achievement tests (e.g., formative tests), plans for feedback of results to students are needed if the tests are to help students improve their learning.

We are more likely to construct appropriate tests and use them effectively if our intentions for use are given careful consideration when teaching plans are made. In constructing a test to measure learning progress, for example, we may want to key each set of items measuring the same learning task to reference materials so that students can do individual study in remedying their learning errors. Thus, by considering the use of test results during instructional planning, we can maximize the contribution that tests make to the teaching-learning process.

6. Achievement tests should provide scores that are relatively free from measurement errors. A well-constructed achievement test should provide consistent results. If we obtain a set of test scores for our students, we need to be confident that we would get similar scores if we tested them at a different time or with a different, but equivalent, sample of items. Unless our measures provide this kind of generalizability concerning student achievement, little confidence can be placed in a given set of scores. The degree to which test scores provide consistent, and thus reliable, results depends on the extent to which errors of measurement are present.

There are a number of factors that increase the amount of error in test scores. Ambiguous test items will introduce errors into the scores. Testing a specific skill with too few test items will provide scores that are more influenced by chance factors than by student performance. The subjectivity of scoring essay tests will introduce errors into the scores. During testing, variations in the students' attention, effort, fatigue, and tendency to guess will influence the consistency of their test responses and thereby introduce error. Many of the sources of measurement error can be controlled through care in the construction and administration of tests. Where the sources of error cannot be eliminated as in scoring essay questions, they can be taken into account during the interpretation of the results. We can make only tentative interpretations and seek verification from other evidence of achievement.

NORM-REFERENCED AND CRITERION-REFERENCED TESTING

As we noted earlier, an achievement test can be used to provide (1) a relative ranking of students, or (2) a description of the learning tasks a student can and cannot perform. Test results of the first type are interpreted in terms of each student's relative standing among other students (for example, "He is third highest in a class of 35 students"). This method of interpreting test performance is called *norm-referenced* interpretation. Test results of the second type are expressed in terms of the specific knowledge and skills each student can demonstrate (for instance, "He can identify all parts of the microscope and demonstrate its proper use"). This method of interpreting test results is called *criterion-referenced* interpretation. Both methods of describing test results are useful. The first tells how an individual's test performance compares with that of others. The second tells in specific performance terms what an individual can do without reference to the performance of others.

Strictly speaking, the terms *norm referenced* and *criterion referenced* refer only to the method of interpreting test results. Thus, both types of interpretation could be applied to the same test. For example, we might say, "Joan surpassed 90 percent of the students (norm-referenced interpretation) by correctly completing 20 of the 25 chemical equations" (criterion-referenced interpretation). The two types of test interpretation are likely to be most meaningful, however, when the test is designed specifically for the type of interpretation to be made. In general, norm-referenced interpretation is facilitated by tests that provide a wide spread of scores so that reliable discriminations can be made among students at various levels of achievement. This is typically done by eliminating easy test items and favoring items of average difficulty. On the other hand, criterion-referenced interpretation is facilitated by including test items that are directly relevant to the learning outcomes, whether or not they are easy for students. Eliminating the easy items would provide incomplete descriptions of student performances because such descriptions would

not include those learning tasks that were mastered by *all* students. Since each type of interpretation is favored by a different approach to test construction, the terms *norm-referenced test* and *criterion-referenced test* have come into widespread use.

A summary of some common characteristics of tests that are specifically designed to emphasize each type of interpretation is presented in Table 1.1. It must be kept in mind, however, that these are primarily matters of emphasis. For example, norm-referenced tests are typically, but not exclusively, used for surveying achievement over a wide range of learning outcomes. By the same token criterion-referenced tests are typically, but not exclusively, used for mastery testing. A review of the characteristics of each testing approach in Table 1.1 will reveal the differences that exist when each test is constructed to serve its principal use.

TABLE 1.1 Summary Comparison of Two Basic Approaches to Achievement Testing

	NORM-REFERENCED TESTING	CRITERION-REFERENCED TESTING
Principal Use	Survey testing.	Mastery testing.
Major Emphasis	Measure individual differences in achievement.	Describe tasks students can perform.
Interpretation of Results	Compare performance to that of other individuals.	Compare performance to a clearly specified achievement domain.
Content Coverage	Typically covers a broad area of achievement.	Typically focuses on a limited set of learning tasks.
Nature of Test Plan	Table of specifications is commonly used.	Detailed domain specifications are favored.
Item Selection Procedures	Items are selected that provide maximum discrimination among individuals (to obtain high score variability). Easy items are typically eliminated from the test.	Includes all items needed to adequately describe performance. No attempt is made to alter item difficulty or to eliminate easy items to increase score variability.
Performance Standards	Level of performance is determined by *relative* position in some known group (ranks fifth in a group of twenty).	Level of performance is commonly determined by *absolute* standards (demonstrates mastery by defining 90 percent of the technical terms).

Constructing Norm-Referenced and Criterion-Referenced Tests

There are more similarities than differences in the preparation of norm-referenced tests (NRT) and criterion-referenced tests (CRT), and, as noted earlier, the differences are largely a matter of emphasis. The following statements highlight some of the major similarities and differences in constructing NRTs and CRTs for measuring achievement.

1. Both typically require specification of the intended learning outcomes as a basis for test construction.
   - NRT—The intended learning outcomes may be described in general or specific terms.
   - CRT—The intended learning outcomes tend to be described in specific terms.
2. Both are typically designed to measure a representative sample of the specified learning outcomes.
   - NRT—Usually covers a broad range of outcomes with few test items per outcome.
   - CRT—Usually covers a delimited domain of outcomes with numerous test items per outcome.
3. Both use a variety of types of test items.
   - NRT—Selection-type items (e.g., multiple choice) are highly favored.
   - CRT—There is somewhat less dependence on selection-type items.
4. Both require the application of a common set of item-writing rules.
   - NRT—The ability of items to *discriminate* among students is emphasized.
   - CRT—The ability of items to *describe* student performance on specific learning tasks is emphasized.
5. Both require attention to the reliability of the results.
   - NRT—The traditional statistical procedures for estimating reliability are appropriate because of high score variability.
   - CRT—The traditional statistical procedures for estimating reliability are inappropriate due to limited score variability (see Chapter 9).
6. Both require care in interpreting the results.
   - NRT—Knowing the nature of the group is especially important in evaluating a student's relative standing.
   - CRT—Knowing the nature of the achievement domain is especially important in describing a student's performance.

These two test types are probably best viewed as the ends of a continuum rather than as discrete categories. Many test publishers have combined the best features of each by making norm-referenced tests more descriptive (e.g., by combining items into meaningful item clusters for C-R interpretation) and by adding norm-referenced interpretation to criterion-referenced tests. This use of dual interpretation requires compromises in test construction and caution in use of the test scores, but the increased information concerning student achievement contributes to the usefulness of the test results.

Most of the discussions of test construction in the following chapters will apply to both test types. Where there are significant differences in test development procedures, these will be noted.

SUMMARY OF POINTS

The emphasis in this chapter can be summarized by the following points.

1. The main purpose of achievement testing is to improve student learning.
2. Achievement tests can be used for measuring entry performance (placement

tests), monitoring learning progress (formative and diagnostic tests), or measuring end-of-instruction achievement (summative tests).

3. Achievement tests can contribute to student motivation, the retention and transfer of learning, student self-evaluation skills, and an evaluation of instructional effectiveness.
4. Minimum competency testing is a nationwide attempt to use tests to improve student learning.
5. The first step in achievement testing is to clearly specify the domain of achievement to be measured.
6. Achievement tests should be designed to measure a representative sample of the instructionally relevant learning tasks specified in the achievement domain.
7. The types of items used in an achievement test should be determined primarily by the learning tasks to be measured.
8. Achievement tests should provide dependable scores that are in harmony with their intended use.
9. Achievement tests can be interpreted by comparing a student's performance to that of others (norm referenced) or by describing the student's performance on a clearly defined set of tasks (criterion referenced).
10. Norm-referenced and criterion-referenced tests are best viewed as the ends of a continuum rather than as discrete categories.
11. Tests can be developed that emphasize norm-referenced interpretation (e.g., survey tests), criterion-referenced interpretation (e.g., mastery tests), or a combination of the two.
12. Dual interpretation requires compromises in test construction and caution in test interpretation but it provides increased information concerning an individual's achievement.

ADDITIONAL READING

BERK, R. A., ED., *A Guide to Criterion-Referenced Test Construction* (Baltimore, Maryland: Johns Hopkins University Press, 1984). See Chapter 1 by A. J. Nitko, "Defining 'Criterion-Referenced Test' " for a review of the many definitions of criterion referencing.

BLOOM, B. S., MADAUS, G. T., AND J. T. HASTINGS, *Evaluation to Improve Learning* (New York: McGraw-Hill, 1981). See Chapters 4, 5, and 6. Provides a comprehensive treatment of summative, diagnostic, and formative evaluation.

GRONLUND, N. E., *Measurement and Evaluation in Teaching*, 5th ed. (New York: Macmillan Publishing Co., Inc., 1985). See Chapter 1. Presents the principles and procedures of classroom testing and its role in the instrucitonal process.

POPHAM, W. J. *Modern Educational Measurement* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1981). See Chapter 2, "Norm-Referenced and Criterion-Referenced Measurement," for a comparison of the two types and examples of educational decisions where each is more useful.

TITLE, C. K., "COMPETENCY TESTING." In *Encyclopedia of Educational Research*, 5th ed. (New York: Macmillan Publishing Co., Inc., 1982), Vol. 1, pp. 333-352. Describes minimum competency testing and the measurement, policy, and legal issues involved.

2

Planning the Test

The key to effective achievement testing is careful planning. It provides greater assurance that our test will measure relevant learning outcomes . . . measure a representative sample of the desired outcomes . . . and provide dependable information on which to base instructional decisions. . . . Test planning involves the identification and specification of precisely what is to be measured.

The planning of an achievement test can take many forms, both professional test makers and classroom teachers have found the following series of steps to be most useful.

1. Determine the purpose of the test.
2. Identify and define the intended learning outcomes.
3. Prepare the test specifications.
4. Construct relevant test items.

It is obvious from these steps that the first consideration in test planning is to determine the type of test to be prepared. This will help clarify what is to be measured and will aid in stating the test specifications in such precise terms that test items can be constructed to call forth the desired performance. If the test planning is carefully done, constructing relevant test items is greatly simplified.

DETERMINING THE PURPOSE OF THE TEST

As we noted in Chapter 1, tests can be used in an instructional program to assess entry behavior (placement test), monitor learning progress (formative test), diagnose learning difficulties (diagnostic test), and measure performance at the end of

instruction (summative test). Each type of test use typically requires some modification in test design. Although the specific makeup of any test depends on the particular situation in which it is to be used, it is possible to identify common characteristics of the various test types. These have been summarized in Table 2.1.

TABLE 2.1 Characteristics of Four Types of Achievement Tests

TYPE OF TEST	FUNCTION OF TEST	SAMPLING CONSIDERATIONS	ITEM CHARACTERISTICS
	Measure prerequisite entry skills	Include each prerequisite entry behavior	Typically, items are easy and criterion-referenced
Placement	Determine entry performance on course objectives	Select representative sample of course objectives	Typically, items have a wide range of difficulty and are norm referenced
Formative	Provide feedback to students and teacher on learning progress	Include all unit objectives if possible (or those most essential)	Items match difficulty of unit objectives and are criterion referenced
Diagnostic	Determine causes of recurring learning difficulties	Include sample of tasks based on common sources of learning error	Typically, items are easy and are used to pinpoint specific causes of error
Summative	Assign grades, or certify mastery, at end of instruction	Select representative sample of course objectives	Typically, items have a wide range of difficulty and are norm referenced

Adapted from P. W. Airasian and G. F. Madaus, "Functional Types of Student Evaluation," *Measurement and Evaluation in Guidance*, 4 (1972), 221-33.

The material in Table 2.1 provides a good general description of the four basic test types we have discussed, but it must be recognized that the categories overlap to some degree. In some instances, a particular test may be designed to serve more than one function. For example, an end-of-unit formative test may be used to provide feedback to students, to pinpoint sources of learning error, and to certify mastery of unit objectives. Similarly, sampling considerations and item characteristics may need to be modified to fit a particular test use or a specific type of instruction. Despite the lack of discrete categories, however, the table highlights the variety of functions that achievement tests can serve and the basic framework for planning achievement tests that are designed to be of maximum usefulness.

IDENTIFYING AND DEFINING THE INTENDED LEARNING OUTCOMES

The learning outcomes measured by a test should faithfully reflect the objectives of instruction. Thus, the first order of business is to identify those instructional objectives that are to be measured by the test and then make certain that they are stated in a manner that is useful for testing. This is easier said than done. It is especially difficult if a clearly defined set of instructional objectives is not available to begin with, as is usually the case. One useful guide for approaching this task is the *Taxonomy of Educational Objectives* (see Bloom and others, 1956). This is a comprehensive system that classifies objectives within each of three domains: (1) cognitive, (2) affective, and (3) psychomotor. The cognitive domain of the taxonomy is concerned with intellectual outcomes, the affective domain with interests and attitudes, and the psychomotor domain with motor skills (see Gronlund, 1985, for a summary of each). Since our concern here is with achievement testing, we shall focus primarily on the cognitive domain.

Cognitive Domain of the Taxonomy

Intellectual outcomes in the cognitive domain are divided into two major classes: (1) knowledge and (2) intellectual abilities and skills. These are further subdivided into six main areas as follows:

KNOWLEDGE

1.00 KNOWLEDGE (Remembering previously learned material)
- 1.10 Knowledge of specifics
  - 1.11 Knowledge of terms
  - 1.12 Knowledge of specific facts
- 1.20 Knowledge of ways and means of dealing with specifics
  - 1.21 Knowledge of conventions
  - 1.22 Knowledge of trends and sequences
  - 1.23 Knowledge of classifications and categories
  - 1.24 Knowledge of criteria
  - 1.25 Knowledge of methodology
- 1.30 Knowledge of the universals and abstractions in a field
  - 1.31 Knowledge of principles and generalizations
  - 1.32 Knowledge of theories and structures

INTELLECTUAL ABILITIES AND SKILLS

2.00 COMPREHENSION (Grasping the meaning of material)
- 2.10 Translation (Converting from one form to another)
- 2.20 Interpretation (Explaining or summarizing material)
- 2.30 Extrapolation (Extending the meaning beyond the data)

3.00 APPLICATION (Using information in concrete situations)

4.00 ANALYSIS (Breaking down material into its parts)
- 4.10 Analysis of elements (Identifying the parts)
- 4.20 Analysis of relationships (Identifying the relationship)
- 4.30 Analysis of organizational principles (Identifying the organization)

5.00 SYNTHESIS (Putting parts together into a whole)
- 5.10 Production of a unique communication
- 5.20 Production of a plan or proposed set of operations
- 5.30 Derivation of a set of abstract relations

6.00 EVALUATION (Judging the value of a thing for a given purpose using definite criteria)

6.10 Judgments in terms of internal evidence
6.20 Judgments in terms of external criteria[1]

As can be seen in this outline the outcomes are arranged in order of increasing complexity. They begin with the relatively simple recall of factual information, proceed to the lowest level of understanding (comprehension), and then advance through the increasingly complex levels of application, analysis, synthesis, and evaluation. The subdivisions within each area are also listed in order of increasing complexity. This scheme for classifying student behavior is thus hierarchical; that is, the more complex behaviors include the simpler behaviors listed in the lower categories.

The cognitive domain of the taxonomy is especially useful in planning the achievement test. It focuses on a comprehensive and apparently complete list of mental processes to be considered when identifying learning outcomes, it provides a standard vocabulary for describing and classifying learning outcomes, and it serves as a guide for stating learning outcomes in terms of specific student performance.

Although the cognitive domain of the taxonomy provides a valuable guide for identifying learning outcomes, not all of the areas listed under this domain will be covered in a particular test or even in a particular course. Moreover, the classification scheme is neutral concerning the relative importance of the learning outcomes listed. Thus, it is the instructor who must decide which learning outcomes will guide their teaching and testing, and how much emphasis each outcome will receive. The taxonomy serves merely as a convenient checklist of outcomes that prevents relevant areas of student performance from being overlooked during the planning of an achievement test.

Stating the General Learning Outcomes

The learning outcomes to be measured by a test are most useful in test construction when they are stated as *terminal performance* that is *observable*. That is, they should clearly indicate the student performance to be demonstrated at the end of the learning experience. The following list of learning outcomes for a unit in the planning of an achievement test illustrates this type of statement. It should be noted that these statements include only objectives that can be tested and that they are stated as *general outcomes*. Before being used for test construction, each one would need to be further defined in terms of specific learning outcomes.

At the end of this unit in achievement test planning the student will demonstrate that he or she:

1. Knows the meaning of common terms.
2. Knows specific facts about test planning.
3. Knows the basic procedures for planning an achievement test.

[1]Reprinted from Benjamin S. Bloom, ed., and others, *Taxonomy of Educational Objectives: Cognitive Domain* (New York: David McKay Co., Inc., 1956), pp. 201-207. Reprinted with permission of the publisher.

4. Comprehends the relevant principles of testing.
5. Applies the principles in test planning.

These statements of general learning outcomes have been deliberately kept free of specific course content so that with only slight modification they can be used with various units of study. As we shall see later, the test specifications provide a means of relating intended outcomes to specific subject matter topics.

This list of general outcomes could, of course, be expanded by making the statements more specific, and in some cases it may be desirable to do so. The number of general learning outcomes to use is somewhat arbitrary, but somewhere between five and fifteen provide a list that is both useful and manageable. Typically, a shorter list is satisfactory for a unit of study, while a more comprehensive list is needed for summative testing at the end of a course.

Defining the General Outcomes in Specific Terms

When a satisfactory list of general learning outcomes has been identified and clearly stated, the next step is to list the specific types of student performance that are to be accepted as evidence that the outcomes have been achieved. For example, what specific types of performance will show that a student "knows the meaning of common terms" or "comprehends the relevant principle of testing"? For these two areas the specific learning outcomes may be listed a follows:

1. Knows the meaning of common terms.
   1.1 Identifies the correct definitions of terms.
   1.2 Identifies the meaning of terms when used in context.
   1.3 Distinguishes between terms on basis of meaning.
   1.4 Selects the most appropriate terms when describing testing procedures.

4. Comprehends the relevant principles of testing.
   4.1 Decribes each principle in his or her own words.
   4.2 Matches a specific example to each principle.
   4.3 Explains the relevance of each principle to the major steps in test planning.
   4.4 Predicts the most probable effect of violating each of the principles.
   4.5 Formulates a test plan that is in harmony with the principles.

Note that the terms used to describe the specific learning outcomes indicate student performance that can be demonstrated to an outside observer. That is, they are *observable* responses that can be called forth by test items. The key terms are listed below to emphasize what is meant by defining learning outcomes in *specific performance terms*.

Identifies	Matches
Distinguishes between	Explains
Selects	Predicts
Describes	Formulates

Action verbs such as these indicate precisely what the student is able to do to demonstrate achievement. Such vague and indefinite terms as "learns," "sees," "realizes," and "is familiar with" should be avoided, since they do not clearly indicate the terminal performance to be measured.

Sample action verbs for stating specific learning outcomes at each level of the cognitive domain of the taxonomy are presented in Table 2.2. Although certain action verbs may be used at several different levels (e.g., "identifies"), the table provides a useful guide for defining intended outcomes in performance terms. For more comprehensive lists of action verbs, see the references by Gronlund (1985, 1985) listed at the end of this chapter.

In defining the general learning outcomes in specific performance terms, it is typically impossible to list all of the relevant types of performance. The proportion that need be listed depends to a large extent on the nature of the test. In planning a test that is to be used to *describe* which learning tasks a student has mastered (criterion-referenced test), we should like as comprehensive a list as possible. For a test that is used to *rank* students in order of achievement (norm-referenced test), however, it is usually satisfactory to include a sufficient number of specific types of performance to clarify what the typical student is like who has achieved the intended outcomes.

TABLE 2.2 Illustrative Action Verbs for Defining Objectives in the Cognitive Domain of the Taxonomy

TAXONOMY CATEGORIES	SAMPLE VERBS FOR STATING SPECIFIC LEARNING OUTCOMES
Knowledge	Identifies, names, defines, describes, lists, matches, selects, outlines
Comprehension	Classifies, explains, summarizes, converts, predicts, distinguishes between
Application	Demonstrates, computes, solves, modifies, arranges, operates, relates
Analysis	Differentiaties, diagrams, estimates, separates, infers, orders, subdivides
Synthesis	Combines, creates, formulates, designs, composes, constructs, rearranges, revises
Evaluation	Judges, criticizes, compares, justifies, concludes, discriminates, supports

PREPARING THE TEST SPECIFICATIONS

The writing of test items should be guided by a carefully prepared set of test specifications. The function of the specifications is to describe the achievement domain being measured and to provide guidelines for obtaining a respresentative sample of test tasks. Although the nature and detail of test specifications can be expected to vary considerably, here we shall describe two of the more commonly recommended procedures. In the construction of norm-referenced achievement tests, one of the most widely used devices has been a two-way chart called a *table of specifications*.

Building a Table of Specifications

Preparing a table of specifications involves (1) selecting the learning outcomes to be tested, (2) outlining the subject matter, and (3) making a two-way chart. The two-way chart describes the sample of items to be included in the test.

Selecting the Learning Outcomes to be Tested. The learning outcomes for a particular course will depend on the specific nature of the course, the objectives attained in previous courses, the philosophy of the school, the special needs of the students, and a host of other local factors that have a bearing on the instructional program. Despite the variation from course to course, most lists of instructional objectives will include learning outcomes in the following areas: (1) knowledge, (2) intellectual abilities and skills, (3) general skills (laboratory, performance, communication, work-study), and (4) attitudes, interests and appreciations. It is in the first two areas covered by the cognitive domain of the taxonomy that achievement testing is most useful. Learning outcomes in the other areas are typically evaluated by rating scales, checklists, anecdotal records, inventories, and similar nontest evaluation procedures. Thus, the first step is to separate from the list of learning outcomes those that are testable by paper-and-pencil test. The selected list of learning outcomes should, of course, be defined in specific terms, as described in the previous section. Clarifying the specific types of performance to be called forth by the test will aid in constructing test terms that are most relevant to the intended learning outcomes.

Outlining the Subject Matter. The stated learning outcomes specify how students are expected to react to the subject matter of a course. Although it is possible to include both the student performance and the specific subject matter the student is to react toward in the same statement, it is usually desirable to list them separately. The reason for this is that the student can react in the same way to many different areas of subject matter, and he or she can react in many different ways to the same area of subject matter. For example, when we state that a student can "define a term in his or her own words," "recall a specific fact," or "identify an example of a principle," these types of performance can be applied to almost any area of subject matter. Similarly, in studying the taxonomy of educational objectives we may expect students merely to recall the categories in it, or we could require them to explain the principles on which it is organized, to summarize its usefulness in test planning, to classify a given set of learning outcomes with it, or to use it in the actual construction of a test. Since particular types of student performance can overlap a variety of subject matter areas, and vice versa, it is more convenient to list each aspect of performance and subject matter separately and then to relate them in the table of specifications.

The content of a course may be outlined in detail for teaching purposes, but for test planning only the major categories need be listed. The following outline of subject matter topics covered in the first two chapters of this book illustrates sufficient detail for the test plan.

A. Role of testing in the instructional process
   1. Instructional decisions and test types
   2. Influence of tests on learning and instruction
B. Principles of achievement testing
   1. Relation to instructional objectives
   2. Representative sampling
   3. Relevance of items to outcomes
   4. Relevance of test to use of results
   5. Reliability of results
   6. Improvement of learning
C. Norm-referenced versus criterion-referenced testing
D. Planning the test
   1. Determining the purpose of the test
   2. Identifying the intended learning outcomes
   3. Preparing the test specifications
   4. Constructing relevant test items

In using the topics in this book for illustrative purposes, there is no implication that the content outline should be limited to the material in a particular book. An achievement test is typically designed to measure all of the course content, including that covered in class discussion, outside reading, and other special assignments. Our example here is meant to illustrate the approximate amount of detail and not the source of the topics to be included.

Making the Two-Way Chart. When the learning outcomes have been selected and clearly defined and the course content outlined, the two-way chart should be prepared. This is called a table of specifications. It relates outcomes to content and indicates the relative weight to be given to each of the various areas. As noted earlier, the purpose of the table is to provide assurance that the test will measure a representative sample of the learning outcomes and the subject matter topics to be measured.

An example of a table of specifications for a norm-referenced summative test on the first two chapters of this book is given in Table 2.3. Note that only the general learning outcomes relevant to these chapters and only the major subject matter categories have been included. A more detailed table may be desirable for test purposes, but this is sufficient for illustration.

The numbers in each cell of the table indicate the number of test items to be devoted to that area. For example, 15 items in the test will measure knowledge of terms; four of them pertain to the "role of tests" in instruction," four to "principles of testing," four to "norm referenced versus criterion referenced," and three to "planning the test." The number of items assigned to each cell is determined by the weight given to each learning outcome and each subject matter area.

A number of factors will enter into assigning relative weights to each learning outcome and each content area. How important is each area in the total learning experience? How much time was devoted to each area during instruction? Which outcomes have the greater retention and transfer value? What relative importance do curriculum specialists assign to each area? These and similar criteria must be con-

TABLE 2.3 Table of Specifications for a Summative Test on Chapters 1 and 2 of This Book

CONTENT \ OUTCOMES	KNOWS			COMPREHENDS	APPLIES	TOTAL NUMBER
	TERMS	FACTS	PROCEDURES	PRINCIPLES	PRINCIPLES	OF ITEMS
Role of Tests in Instruction	4	4		2		10
Principles of Testing	4	3	2	6	5	20
Norm Referenced versus Criterion Referenced	4	3	3			10
Planning the Test	3	5	5	2	5	20
Total Number of Items	15	15	10	10	10	60

sidered. In the final analysis, however, the weights assigned in the table should faithfully reflect the emphasis given during instruction. In Table 2.3, for example, it is assumed that twice as much emphasis was given to "planning the test" (20 items) as was given to "norm referenced versus criterion referenced" (10 items). Similarly, it is assumed that knowledge outcomes were given approximately two-thirds of the emphasis during instruction (40 items) and that comprehension and application outcomes were each given approximately one-sixth of the total emphasis (10 items each).

In summary, preparing a table of specifications includes the following steps:

1. Identify the learning outcomes and content areas to be measured by the test.
2. Weigh the learning outcomes and content areas in terms of their relative importance.
3. Build the table in accordance with these relative weights by distributing the test items proportionately among the relevant cells of the table.

The resulting two-way table indicates the type of test needed to measure the learning outcomes and course content in a balanced manner. Thus, the table of specifications serves the test maker like a blueprint. It specifies the number and the nature of the items in the test, and it thereby provides a guide for item writing.

Specifications for Criterion-Referenced Tests

Whether a table of specifications is useful in preparing a criterion-referenced test depends greatly on the nature of the achievement domain being measured. In some cases the test may cover such a limited area that a table of specifications is unnecessary. In constructing a test to measure "knowledge of the categories in the cognitive domain of the taxonomy," for example, a list of specific ways students are to demonstrate their knowledge may provide a sufficient basis for test planning. The specifications could then simply indicate the number of items to be used to measure each specific type of performance as shown in Table 2.4. For some purposes it may be desirable to use a listing like that in Table 2.4 *and* a table of specifications. In any event the test plan should indicate the achievement domain to be measured and the specific nature of the student performance to be accepted as evidence of achievement.

TABLE 2.4 Specific Knowledge Outcomes to Result from Study of the Cognitive Domain of the *Taxonomy* and the Number of Test Items for Each Intended Outcome

TYPES OF STUDENT PERFORMANCE	NUMBER OF ITEMS
Describes the meaning of each major category.	6
Identifies the hierarchical order of the major categories.	4
Identifies a verb that best fits each main category.	6
States a specific learning outcome for each main subcategory.	10
Distinguishes between given pairs of categories.	4

When criterion-referenced tests are being prepared by more than one person (e.g., for a mastery testing program) more detailed test specifications are desirable. These specifications should include a description of the test items to be used plus a sample item that illustrates the item characteristics. A brief example of such specifications is shown in Figure 2.1.

An even more comprehensive and detailed set of specifications for criterion-referenced tests has been recommended by Popham (1978, 1984). He suggests writing separate test specifications for each set of items that measure the same class of student performance. This typically involves a clearly defined and delimited achievement domain and includes the following test specification components:[2]

1. *General description:* A brief description, in general terms, of the behavior being assessed by the test.
2. *Sample item:* An illustrative item that reflects the test item attributes to be delimited in the following two components.
3. *Stimulus attributes:* A series of statements that attempt to delimit the class of stimulus material that will be encountered by the examinee.
4. *Response attributes:* A series of statements that attempts either to (a) delimit the class of response or response options from which the student makes *selected responses* or (b) explicate the standards by which an examinee's *constructed responses* will be judged.
5. *Specification supplement:* In certain cases it may be necessary to add an appendix or supplement to the preceding four components. This supplement typically provides a more detailed listing or explanation of eligible content.

FIGURE 2.1. Specifications for a Ten-Item Test Measuring One Specific Comprehension Outcome.

General Outcome:	Comprehends the meaning of measurement terms.
Specific Outcome:	Identifies an example of each term.
Type of Test:	Multiple choice (10 items).
Item Characteristics:	Each test item will consist of a question or incomplete statement followed by four possible answers placed in alphabetical order. The *correct* answer will be an example that best fits the meaning of the term. The *incorrect* alternatives (distracters) will be made plausible by using common student misunderstandings and by matching the correct answer in terms of length, content, and grammatical structure.
Sample Item:	Which of the following is an example of a *performance term*? *A Defines B Fears C Realizes D Thinks

[2]W. J. Popham, *Criterion-Referenced Measurement* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1978), pp. 121-122.

Such detailed specifications may run several pages and thus be time consuming to prepare (see Popham, 1978, 1984, for examples). If carefully done, however, the specifications will clearly indicate the nature of the student performance being measured and the specific characteristics of the test items to be prepared. This procedure is especially useful where pools of test items are being prepared for school-wide use and for storage in computer item banks. In this case, the specifications provide the needed guidelines for preparing pools of instructionally relevant and functionally similar test items. The detailed descriptions are, of course, also useful in communicating to the test user specifically what the test measured and, thus, what the test scores represent. Because criterion-referenced tests are designed to describe the tasks that students can perform, detailed specifications are needed for proper interpretation.

CONSIDERATIONS IN CONSTRUCTING RELEVANT TEST ITEMS

The construction of a set of relevant test items is greatly simplified if the intended learning outcomes have been clearly defined and the test specifications carefully prepared. The quality of the test will then depend on how closely the test maker can match the specifications. Here we shall confine our discussion to some of the general specifications in preparing test items. More detailed procedures and rules for item writing will be described in the chapters that follow.

Selecting the Types of Test Items to Use

The items used in achievement tests can be classified as either *selection-type* items or *supply-type* items. The selection-type item presents students with a set of possible responses from which they are to select the most appropriate answer. The supply-type item requires students to create and supply their own answers. These two major categories can be used to classify the most widely used item types as follows.

SELECTION-TYPE ITEMS

1. Multiple choice
2. True-false
3. Matching
4. Interpretive exercise

SUPPLY-TYPE ITEMS

1. Short-answer
2. Essay (Restricted response)
3. Essay (Extended response)

These categories are sometimes referred to as *recognition* and *recall* items. This is a case of mislabeling that confuses the method of responding with the mental

reaction needed to make the response. When measuring knowledge of facts, test responses might be limited to either the recognition or recall of the answer. However, when measuring complex learning outcomes with selection-type items, the answer is not achieved through mere recognition of a previously learned answer. It typically involves some use of higher mental processes to arrive at a solution (e.g., verbal or mathematical reasoning) before the correct answer can be selected. Similarly, a short-answer item may require reasoning or problem solving rather than simply recalling and supplying factual information. Essay answers, of course, typically require analysis, synthesis, and evaluation skills in addition to recall. Using the *selection* and *supply* labels makes clear how the responses are made but it does not imply limits on the types of learning outcomes that can be measured with each.

In deciding which item types to use in a test, a guiding principle should be: *Use the item types that provide the most direct measures of student performance specified by the intended learning outcome.* Thus, if you want to determine if students can spell, have them spell from dictation. If you want to determine if students can solve mathematics problems, have them solve problems and supply the answers. If you want to determine if students can write, have them write something. Use selection-type items for supply-type outcomes only if there is a compelling reason for doing so (e.g., electronic scoring) and then take into account, during interpretation of the results, that a less direct measure has been used. In some cases, of course, both types of items are useful in the same area. For example, a writing project may provide the best evidence of writing skill but a selection-type test would provide the most systematic coverage of the elements of grammar needed for effective writing.

There are a number of achievement areas where either selection-type items or supply-type items would measure equally well. In these cases, our choice between them must be based on other item characteristics. The preparation of good selection-type items is difficult and students can get a proportion of answers correct by guessing. However, these disadvantages are offset by the fact that (1) they can be scored quickly and objectively (i.e., scorers agree on the answers), (2) they eliminate bluffing, (3) they eliminate the influence of writing skill, (4) they provide an extensive sample of student performance (because of the large number of items used), and (5) they provide for the identification of specific learning errors. In comparison, supply-type items are easier to construct (although harder than commonly believed) but more difficult to score. The scoring of short-answer items is contaminated by answers of varying degrees of correctness and by adjustments needed for misspellings. The scoring of essay tests is tedious, time consuming and influenced by bluffing, writing skill, and the shifting of standards during scoring. Another major shortcoming of supply-type items is the limited sample of learning tasks that can be measured. The short-answer item is restricted primarily to measuring knowledge outcomes. Although the essay test is especially suited to measuring complex learning outcomes, its sampling is limited by the relatively few questions that can be included in a test.

A summary comparison of the relative merits of selection-type and supply-type items is presented in Table. 2.5.

TABLE 2.5 Summary of the Relative Merits of Selection-Type Items and Supply-Type Items

CHARACTERISTIC	SELECTION-TYPE ITEMS	SUPPLY-TYPE ITEMS: SHORT-ANSWER	ESSAY
Measures factual information	Yes	Yes	Yes(*)
Measures understanding	Yes	No(**)	Yes
Measures synthesis	No(**)	No(**)	Yes
Easy to construct	No	Yes	Yes
Samples broadly	Yes	Yes	No
Eliminates bluffing	Yes	No	No
Eliminates writing skill	Yes	No	No
Eliminates blind guessing	No	Yes	Yes
Easy to score	Yes	No	No
Scoring is objective	Yes	No	No
Pinpoints learning errors	Yes	Yes	No
Encourages originality	No	No	Yes

(*)The essay test can measure knowledge of facts, but because of scoring and sampling problems it probably should not be used for this purpose.

(**)These items can be designed to measure limited aspects of these characteristics.

Although this should serve as a guide in selecting the types of items to use in a given test, as noted earlier, the most important question to ask is: *Does this item type provide the most direct measure of the intended learning outcome?* If the various item types are equal in this regard, the selection-type items would be favored because of the broad sampling, the objective scoring, and the pinpointing of specific learning errors.

Matching Items to Specific Learning Outcomes

Effective achievement testing requires that a set of test items be constructed that calls forth the performance described in the intended learning outcomes. While we can never be certain of a perfect correspondence between outcome and item, the following examples illustrate how items should be written to measure the specific type of performance stated in the specific learning outcome.

EXAMPLES

Specific Learning Outcome: Defines terms in student's own words.

Directions: Define each of the following terms in a sentence or two.

1. Taxonomy
2. Cognitive
3. Measurement
2. Evaluation

Specific Learning Outcome: Identifies procedural steps in planning for a test.

1. Which one of the following steps should be completed first in planning for an achievement test?[3]

[3]The correct answer is indicated throughout this book by an asterisk.

A. Select the types of test items to use.
B. Decide on the length of the test.
*C. Define the intended learning outcomes.
D. Prepare the test specifications.

Specific Learning Outcome: Identifies the hierarchical order of the categories in the cognitive domain of the taxonomy.

1. Which one of the following categories in the taxonomy indicates the highest level of learning?
   A. Analysis
   B. Application
   C. Comprehension
   *D. Synthesis

Specific Learning Outcome: Distinguishes between sound and unsound principles of achievement testing.

Directions: Read each of the following statements. If the statement indicates a sound principle of achievement testing, circle the S; if it indicates an unsound principle, circle the U.

*S U 1. The specific learning outcomes to be tested should be stated in terms of student performance.
S *U 2. Achievement testing should be limited to outcomes that can be measured objectively.
*S U 3. Each achievement test item should measure a clearly defined type of student performance.

Specific Learning Outcome: Identifies examples of properly stated learning outcomes.

1. Which one of the following learning outcomes is properly stated in performance terms?
   A. Student realizes the importance of tests in teaching.
   B. Student has acquired the basic principles of achievement testing.
   C. Student demonstrates a desire for more experience in test construction.
   *D. Student predicts the most probable effect of violating a test construction principle.

It should be noted in these examples that each specific learning outcome provides a precise definition of the student performance to be measured, and the test item simply provides a task that makes measurement of the specified performance possible.

Improving the Functioning Content of Items

If test items are to call forth the performance described in the intended learning outcomes, great care must be taken in phrasing the items. We need to eliminate all barriers that might prevent a knowledgeable person from responding and all clues that might lead the uninformed to the correct answer. Only those who have achieved the outcome being measured should get the item right. All others (no matter how intelligent) should miss it.

Some of the common *barriers* to be avoided during test preparation are:

Vocabulary that is unnecessarily difficult.
Sentence structure that is unnecessarily complex.
Statements containing ambiguity.
Statements that are excessively wordy.
Unclear pictorial materials.
Directions that are vague.
Material reflecting race, ethnic, or sex bias.

Awareness of such barriers during the planning and preparation of the test is the first step in their elimination. Essentially we can avoid these barriers by (1) writing each test item so that it presents a clearly formulated task, (2) stating the items in simple, clear language, (3) keeping the items free from biased and nonfunctional material, and (4) using a test format and directions that contribute to effective test taking. Much of the material presented later in this book is directed toward constructing tests that prevent extraneous factors from distorting the test results.

Some of the common *clues* to be avoided during test preparation are:

Verbal associations that give away the answer.
Grammatical inconsistencies that eliminate wrong answers.
Specific determiners that make certain answers probable (e.g., sometimes) and others improbable (e.g., always).
Stereotyped or textbook phrasing of correct answers.
Length or location of correct answers.
Material in an item that aids in answering another item.

Just as in controlling extraneous factors that provide barriers to the correct answer, clues such as these can be eliminated by being aware of them and by following sound principles of test construction. Many of the rules for item writing and test preparation in the chapters that follow provide guidelines for this purpose.

Selecting the Proper Item Difficulty

Decisions concerning item difficulty are guided to a large extent by the nature of the achievement test being prepared. If the test is to be criterion referenced, item difficulty is determined by the difficulty of the learning tasks described in the specific learning outcomes. If the test is to be norm referenced, item difficulty is deliberately altered to obtain a wide spread of test scores. These different approaches to item difficulty constitute one of the major distinctions between criterion referenced and norm referenced testing.

Because a criterion referenced test is designed to describe the specific learning tasks an individual can and cannot perform, item difficulty should match the difficulty of the task. If the task is easy, the test items should be easy. If the task is difficult, the test items should be difficult. No item should be eliminated simply because most students might be expected to answer it correctly or because it might be answered incorrectly by most students. Likewise no attempt should be made to alter item difficulty to obtain a spread of test scores. What we seek in a criterion referenced test is a set of test items that can be used to describe how well a student performs on a clearly defined domain of learning tasks, without reference to the performance of others. To serve this function effectively, the test items must match the learning tasks as closely as possible in all respects, including item difficulty.

Norm referenced achievement tests are designed to rank individuals in order of their achievement. For this purpose a wide spread of test scores is desired so that a dependable (that is, reliable) ranking is obtained. For example, we can say with greater confidence that Mary has achieved more than Tom if the difference in test scores is ten points rather than two. Thus, the ability of test items to discriminate among students is vital to norm-referenced testing, and typically the greater the spread of scores, the better.

The desired score variability in norm-referenced tests is obtained by eliminating the very easy items (those likely to be answered correctly by all students) and by constructing the majority of items at an average level of difficulty, that is, a level at which approximately one half of the students answer correctly. Although some easy items may be desirable at the beginning of the test for motivational purposes and some difficult ones at the end to challenge the high achievers, items of average difficulty should be favored because they provide for maximum discrimination among individuals. This will be described in greater detail in Chapter 7, where item analysis is considered.

In deliberately altering item difficulty in norm-referenced tests to provide for the desired spread of scores, care should be taken to keep the items relevant to the learning outcomes to be measured. In measuring the ability to distinguish between concepts, for example, item difficulty can be increased by calling for finer discriminations. Similarly, in measuring the ability to apply principles to new situations, items can be constructed that call for increasingly complex applications. Item difficulty should not be increased by measuring more obscure material or by overloading the test with items on a particular learning outcome that happens to be difficult. Although a norm-referenced achievement test is designed to rank students from high to low, that ranking should represent the relative degree to which the instructional objectives are being achieved. If irrelevant difficulty is introduced, the items will provide a less valid measure of the intended learning outcomes.

Determining the Number of Items

The test specifications should indicate what sample of items is to be included in the test and thus the number of items to be used. Although the number of items is determined primarily by the purpose of the test and the types of test items to be prepared, there are practical constraints that must also be considered. Two major ones are the age of the students tested and the time available for testing. In the testing of elementary school students, the testing time typically should be no more

than 30 minutes so that proper motivation is maintained. At the high school and college levels, students can be given tests lasting several hours (for example, a final examination), but most tests are limited to a testing time of 40 to 50 minutes because of the length of the typical class period.

In matching the number of items to available testing time, we are faced with the problem of estimating how many items students can complete per minute. Unfortunately there are no simple answers. It depends on the type of test item, the complexity of the learning outcome measured, and the age of the students. As a rule of thumb, high school and college students should be able to answer one multiple-choice item, three short-answer items, or three true-false items per minute when the items are measuring knowledge outcomes. For measuring more complex learning outcomes such as comprehension and application, and for testing younger age groups, more time per item is needed. In estimating the number of items to be used, keep in mind the slower students in the group, for it is desirable to give all students an opportunity to complete the test. Experience in testing a given group of students is frequently the only dependable guide for determining proper test length.

In addition to our concern with total test length, consideration must be given to the number of test items needed for each type of interpretation to be made. This issue is especially crucial in criterion-referenced interpretation, where we want to describe student performance in terms of each intended learning outcome. For this purpose we should use at least ten items per outcome. Where practical constraints make it necessary to use fewer than ten items for an intended outcome, only tentative judgments should be made and these should be verified by other means. In some cases it is possible to combine items into larger item clusters for a more meaningful interpretation.

TECHNICAL CONSIDERATIONS IN TEST PLANNING

The two most important characteristics of a well-constructed achievement test are *validity* and *reliability*. These characteristics should be of prime concern during test construction and most of the suggestions in this book are directed toward preparing tests that provide valid and reliable scores.

Validity refers to the degree to which test scores serve their intended use. In the case of achievement testing, we intend to use the scores to describe the extent to which students have achieved the intended learning outcomes. Since we can measure only a sample of the student performance that represents satisfactory achievement of the outcomes, the validity of our scores depends mainly on the adequacy of the test sample. Ideally, our test should provide scores that are based on a representative sample of instructionally relevant tasks. This content-related evidence of validity can best be obtained by following a systematic procedure during test development. This includes (1) clearly defining the domain of learning tasks to be measured, (2) carefully preparing the test specifications, and (3) constructing a representative sample of relevant test items. Unless these procedures are followed, the validity of the test scores will be in doubt.

Reliability refers to the *consistency* of measurement. For example, if a student earns a score of 60 on a test, we would like to be able to say that the 60 accurately represents the student's test performance. Thus, if we tested the student at a different time or with a different sample of equivalent items, we would expect to obtain a similar score. This consistency of measurement would indicate that the score is relatively free from errors of measurement and thus one we can depend on (i.e., it has "rely-ability").

We cannot, of course, expect test performance to be perfectly consistent over different occasions or over different samples of the same achievement domain. Such factors as ambiguities in test items, fluctuations in attention, and luck in guessing can introduce errors that cause test scores to vary. An important goal in testing is to keep these errors of measurement to a minimum so that our test scores are as reliable as possible.

In addition to being important in its own right, reliability is necessary to obtain validity. After all, if an individual's test score fluctuated widely on a given sample of items, we could not expect the test to provide a valid measure of achievement. Thus, *reliability provides the consistency that makes validity possible*. It should be noted, of course, that consistency of results is just one important requirement for validity. We could be consistently measuring the wrong thing or using the scores inappropriately. Thus, reliability is a necessary, but not a sufficient, condition for validity.

Both the validity and the reliability of a test can be "built in" during test construction. When we select a representative sample of learning tasks to include in our test, we are providing for validity. When we include an adequate number of items in our test, we are providing for reliability. When we construct relevant test items that are free from ambiguity, clues, and other technical defects, we are providing for both improved validity and reliability. In fact, most of the suggestions for constructing achievement tests presented in this book are directed toward improving the validity and reliability of our test scores.

A more detailed discussion of validity and reliability and methods for determining them is presented in Chapter 9. Although high quality achievement tests can be constructed without this technical background, a fuller understanding of these basic concepts contributes to improved skill in test construction and test interpretation.

OTHER CONSIDERATIONS IN TEST PLANNING

In this chapter, we have emphasized those aspects of test planning that are concerned with the preparation of an achievement test that measures a balanced sample of clearly defined learning outcomes. A complete test plan will, of course, also consider such things as test directions, arrangement of the items in the test, scoring, and whether to correct for guessing. These and similar factors will be discussed in Chapter 7, after we have described the procedures for writing the various types of items that are typically used in achievement tests.

SUMMARY OF POINTS

The emphasis in this chapter can be summarized by the following points.

1. Test planning should be guided by the nature of the learning tasks to be measured and the use to be made of the results.
2. Test planning should include defining the intended learning outcomes in terms of student performance.
3. Test specifications should be prepared before writing or selecting test items.
4. Test specifications can vary from a two-fold table to a detailed description of the achievement to be measured and the nature of the test to be used.
5. The types of items used in a test should be determined by how directly they measure the intended learning outcomes and how effective they are as measuring instruments.
6. Each test item should provide a task that matches the student performance described in a specific learning outcome.
7. The functioning content of test items can be improved by eliminating irrelevant barriers and unintended clues during item writing.
8. For criterion-referenced interpretation, the difficulty of a test item should match the difficulty of the learning task to be measured.
9. For norm-referenced interpretation, item difficulty may be altered to provide a larger spread of scores, but care must be taken not to introduce irrelevant difficulty (e.g., by using obscure material.)
10. An achievement test should be short enough to permit all students to attempt all items during the testing time available.
11. A test should contain a sufficient number of test items for each type of interpretation to be made. Interpretations based on fewer than ten items should be considered highly tentative.
12. Validity and reliability are the two most important characteristics of achievement testing.
13. An achievement test will provide valid and reliable results if it measures a representative sample of instructionally relevant tasks and provides scores that are relatively free of measurement errors.
14. Validity and reliability are "built in" during test construction and, thus, should be considered during all stages of test development.

ADDITIONAL READING

BERK, R. A., ED., *A Guide to Criterion-Referenced Test Construction* (Baltimore, Maryland: Johns Hopkins University Press, 1984). See Chapter 2 by W. J. Popham, "Specifying the Domain of Content or Behaviors" for a description and illustration of criterion-referenced test specifications.

BLOOM, B.S., ED., *et al., Taxonomy of Educational Objectives: Cognitive Domain* (New York: David McKay Co., Inc., 1956). Describes and illustrates the categories in the cognitive domain.

GRONLUND, N.E., *Measurement and Evaluation in Teaching*, 5th ed. (New York: Macmillan Publishing Co., Inc., 1985). See Chapter 2, "Preparing Instructional Objectives" and Chapter 5, "Planning the Classroom Test" for a comprehensive description of how to define intended learning outcomes and their role in planning classroom achievement tests. See the appendix for a description of the cognitive, affective, and psychomotor taxonomies.

GRONLUND, N. E., *Stating Objectives for Classroom Instruction*, 3rd ed. (New York: Macmillan Publishing Co., Inc., 1985). Describes and illustrates how to state instructional objectives as intended learning outcomes. Includes a summary of the cognitive, affective, and psychomotor taxonomies with illustrative objectives and a list of performance terms for each of the categories.

POPHAM, W. J., *Criterion-Referenced Measurement* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1978). See Chapter 6, "Preparing Criterion-Referenced Test Specifications" for illustrative specifications for criterion-referenced tests.

ROID, G., AND T. HALADYNA, *A Technology of Test-Item Writing* (New York: Academic Press, 1981). Describes new methods for writing criterion-referenced test items.