CHAPTER

14

ASSEMBLING, ADMINISTERING, AND APPRAISING CLASSROOM TESTS AND ASSESSMENTS

Care in preparing an assessment plan and constructing relevant test items and assessment tasks should be followed by similar care in reviewing and editing the items and tasks, preparing clear directions, and administering and appraising the results. Classroom assessments also can be improved by using simple methods to analyze student responses and by building a file of effective items and tasks.

Effective classroom assessment begins with an assessment plan that specifically describes the instructional objectives and content to be measured and the relative emphasis to be given to each intended learning outcome. This is followed by the selection of the most appropriate item and task formats (e.g., multiple choice, essay, or hands-on performance assessment) and the preparation of items and tasks that are relevant to the learning outcomes specified in the plan. These steps have received considerable attention in the preceding chapters because they are crucial to the validity of the assessment. The only way we can ensure that a classroom test will serve its intended purpose is to identify the learning outcomes we want to measure and then to construct test items and assessment tasks that call forth the specific performance described in the learning outcomes. We must also assemble the items and tasks, prepare directions, administer the instrument, score the student responses, and interpret and appraise the results.

Our goal throughout the preparation and use of classroom tests and assessments is to obtain valid evidence of student learning. In the final analysis, valid measurement of achievement is the product of a systematically controlled series of steps, beginning with the identification of instructional objectives and ending with the scoring and interpretation of results. Although validity is built in during the construction of the items and tasks, systematic procedures of assembly, administration, and scoring will provide greater

assurance that the items and tasks will function as intended. Appraising the test items and assessment tasks after they have been administered can also help improve their quality. Procedures for analyzing student responses provide information for evaluating the effectiveness of each item or task and for detecting weaknesses that should be corrected. This information is useful when reviewing the results with students, and it is indispensable when building a file of high-quality items and tasks for future use.

Performance-based assessments typically involve a relatively small number of tasks because each task requires a substantial amount of time. Thus, some of the issues addressed in this chapter apply more to classroom tests that involve a substantial number of items and/or restricted response tasks. For example, if a single extended-response, performance-based task is to be administered, the issues of test assembly and some aspects of item analysis are not relevant. Hence, in this chapter we focus on classroom tests with a substantial number of items and make note where the same or related considerations apply to performance-based assessments.

ASSEMBLING THE CLASSROOM TEST

The preparation of items for use in a test is greatly facilitated if the items are properly recorded, if they are written at least several days before they are to be used, and if extra items are constructed.

Recording Test Items

When constructing the test items, it is desirable to write each one in a form that makes it possible to easily address and retrieve individual items. The items may be written into a word-processing program or onto index cards. In addition to the test item, the item record should contain information concerning the instructional objective, the specific learning outcome, and the content measured by the item. A space should also be reserved for item-analysis information to allow room to record the data each time the item is used. An example of this type of informational record is presented later in the chapter in relationship to the discussion of item analysis (Figure 14.2).

Item records, whether in the form of separate sheets in a standard word-processing system or physical cards, provide flexibility. As the items are reviewed and edited, they can be eliminated, added, or revised with little difficulty. The same holds true when arranging the items for the test: They can be arranged and rearranged merely by sorting the records. The flexibility of this recording system also makes it easy to add the items to a computer item bank. Specialized software for item writing and item banks is available, but functions of maintaining a bank of items can also be readily handled by a standard word-processing system.

Reviewing Test Items and Assessment Tasks

No matter how carefully items or tasks have been prepared, defects inadvertently creep in during construction. As we concentrate on the clarity and conciseness of a question, a

verbal clue slips in unnoticed. As we attempt to construct more challenging items, we unwittingly introduce some ambiguity. As we rework a multiple-choice item to make the incorrect choices more plausible, the behavior called forth by the item is unintentionally modified. As we attempt to increase the authenticity of a task for problems faced outside the classroom, we introduce unintended reliance on access to resources that put students from families with limited resources at an unfair disadvantage. In short, we focus so closely on some aspects of item or task construction that we overlook others. This results in an accumulation of unwanted errors that may distort the function of the item or task. Such technical defects can be most easily detected by (a) reviewing the items and tasks after they have been set aside for a few days, and (b) asking a fellow teacher to review and criticize them.

In reviewing test items and tasks, we should try to view them from the student's perspective as well as from that of the teacher. From these two vantage points, each item or task should be read carefully and its effectiveness judged. The following questions will help you analyze the quality of each item or task.

1. **Is the format appropriate for the learning outcome being measured?** If the learning outcome calls for the definition of a term, for example, then a supply-type item (e.g., short-answer item) would be appropriate and a selection-type item (e.g., multiple choice) would be clearly inappropriate. If the learning outcome calls for the ability to collect, organize, integrate, and present information in the form of a coherent argument, then nothing short of a performance-based task will suffice. On the other hand, if the intended outcome was simply the identification of the correct definition, then a selection-type item would be adequate. Thus, the first step is to check whether the format is suitable for the type of student performance described in the testing and assessment plan. The action verb in the statement of each specific learning outcome (e.g., *defines, describes, identifies*) indicates which item format is more appropriate.

2. **Does the knowledge, understanding, or thinking skill called forth by the item or task match the specific learning outcome and subject-matter content being measured?** When a table of specifications has been used as a basis for constructing items and tasks, this is merely a matter of checking to see whether the item or task is still relevant to the same cell in the table. If, for example, an item's functioning content has shifted during construction, the item should be either modified so that it serves its original purpose or reclassified in light of the new purpose. In any case, the response called forth by an item or task should agree with the purpose for which it is to be used.

3. **Is the point of the item or task clear?** A careful review of items and tasks often reveals ambiguity, inappropriate vocabulary, and awkward sentence structure that were overlooked during their construction. Returning to items and tasks after they have been set aside for a few days provides a fresh outlook that makes such defects more apparent. The difficulty of the vocabulary and the complexity of the sentence structure must, of course, be judged in terms of the students' maturity level. At all levels, however, ambiguity should be removed. In its final form, each item or task should be so clearly worded that all students understand its meaning. The quality of student responses should be determined solely by whether they possess the knowledge, understanding, or skill being measured.

4. **Is the item or task free from excessive verbiage?** Often, items become excessively wordy because of awkward sentence structure or the inclusion of nonfunctional material. Some teachers justify the use of an item by including a statement or two concerning the problem's importance. Others expand a simple question into an elaborate story situation to make the item more interesting. Although adding such nonfunctional material may be useful in some instances, items and tasks are generally more effective when the problem is stated as concisely as possible. When reviewing items, the content of each item should be analyzed to determine the functional elements leading to the correct response. If there are any elements that the students may disregard entirely and still respond correctly, they probably should be removed. See the "Guidelines" box.

5. **Does the item have an answer that would be agreed on by experts? How well would experts agree about the degree of excellence of task performances?** This is seldom a problem with factual material, which usually can be judged as correct or incorrect. It is more of a problem with selection-type items that ask for the best reason, the best method, or the best interpretation. The problem is greatest with tasks requiring extended performances where qualified judges may differ in their evaluation of performances. If experts agree on the best response, fine, but do not include items that require students to endorse someone's unsupported opinion (even if it happens to be yours), and do not evaluate performances on tasks simply in terms of your personal preferences.

6. **Is the item or task free from technical errors and irrelevant clues?** The checklists for reviewing each of the item types, presented in Chapters 7 through 11, list points to consider in searching out technical errors and irrelevant clues. As noted earlier, an irrelevant clue for a selection-type item is any element that leads the poor achiever to the correct answer and thereby prevents the item from functioning as intended. These include (a) grammatical inconsistencies, (b) verbal associations, (c) specific determiners (e.g., words such as *always* and *never*), and (d) some mechanical features, such as correct statements tending to be longer than incorrect ones. Most of these clues can be removed merely by trying to detect them during the item review. They somehow seem more obvious after the items have been set aside for a while.

7. **Is the item or task free from racial, ethnic, and gender bias?** A final check should be made to make certain that the vocabulary and problem situation in each item or task would be acceptable to the members of all groups and would have a similar meaning to them. An effort should be made to remove any type of stereotyping, such as always portraying minorities in subservient roles, women in homemaking roles, and the like. A judicious and balanced use of different roles for minorities and males and females should contribute to more effective assessment.

When possible, it can be useful to get fellow teachers to review your test items and assessment tasks. With right–wrong or best-answer items, they should be asked to read each item, indicate the answer, and note any technical defects. If an answer does not agree with the key, it may be because the question is ambiguous. Asking another teacher to "think out loud" when deciding on the answer will usually reveal the misinterpretation of the question and the source of the ambiguity. For performance tasks requiring extended

GUIDELINES

Reviewing and Revising Test Items

1. Matching the learning outcome.
Specific Learning Outcome: Identifies the use of weather instruments.
Item: Describe how the hygrometer works.
Improved: The hygrometer is used to measure
A. Air pressure.
*B. Humidity.
C. Rainfall.
D. Wind velocity.

2. Clarifying the point of the item and the desired response.
Item: Earthquakes are detected by _____ .
Improved: Earthquakes are detected by an instrument called a(n) _____ (seismograph).

3. Removing excessive verbiage from multiple-choice stems.
Item: In which one of the following regions of the United States can we expect annual rainfall to be the greatest?
Improved: In which region of the Uniited States is yearly rainfall greatest?
A. Midwest
B. New England
*C. Pacific Northwest
D. Southwest

4. Removing excessive verbiage from multiple-choice alternatives.
Item: In which direction do tornadoes move?
*A. They move toward the Northeast.
B. They move toward the Northwest.
C. They move toward the Southeast.
D. They move toward the Southwest.
Improved: Tornadoes move toward the
*A. Northeast.
B. Northwest.
C. Southeast.
D. Southwest.

5. Keeping the reading level low.
Item: *T F There is a dearth of information concerning the possibility that life exists on Mars.
Improved: *T F There is a lack of information concerning life on Mars.

6. Removing verbal clues.
Item: Evaporation is shown by water changing to
A. Dew.
B. Ice.
*C. Water vapor.
Improved: Evaporation is shown by water changing to
A. Dew.
B. Ice.
*C. Steam.

responses, reviewers identify the nature of the performances that they would expect and the qualities they would look for in evaluating the performances as well as possible defects in the task itself. This is how other persons can be most useful. Reviewers will be less helpful in evaluating the types of responses called forth by the items because this requires a knowledge of what the students have been taught. Only the teacher who prepared the item or task knows for sure whether it is likely to measure understanding or merely the recall of a previously learned response.

When items or tasks have been revised and those to be included in the test or assessment have been tentatively selected, ask the following questions.

1. Does the set of items and tasks measure a representative sample of the learning outcomes and course content included in the assessment plan?
2. Are there enough items or tasks for each interpretation to be made?
3. Is the difficulty of the items and tasks appropriate for the measurement purpose and for the students for whom the test or assessment is intended?
4. Are the test items free from overlapping so that the information in one does not provide a clue to the answer in another?

The first question can be answered by comparing the final selection of items and tasks with the table of specifications or other assessment plan. Answers to the last three are determined by reviewing the items and tasks in each content area and as a total set. Affirmative answers to these questions mean the items and tasks are ready to be assembled for administration. See the box "Review of Test Items Selected from Item Banks" for guidelines on selecting items from published sources.

Arranging Items in the Test

There are various methods of grouping items in an achievement test, and the method will vary somewhat depending on the use of the results. For most classroom purposes, the items can be arranged by a systematic consideration of (a) the types of items used, (b) the learning outcomes measured, (c) the difficulty of the items, and (d) the subject matter measured.

First, the items should be arranged in sections by item type. That is, all true–false items should be grouped together, then all matching items, then all multiple-choice items, and so on. This arrangement requires the fewest sets of directions, it is the easiest for the students because they can retain the same mental set throughout each section, and it greatly facilitates scoring. When two or more item types are included in a test, there is also some advantage in keeping the simpler item types together and placing the more complex ones in the test, as follows:

1. True–false or alternative-response items
2. Matching items
3. Short-answer items
4. Multiple-choice items
5. Interpretive exercises
6. Restricted-response essay questions
7. Restricted-response performance tasks

Extended-response essay questions and performance tasks usually take enough time that they would be administered alone. If combined with some of the other types of items and tasks listed previously, the extended-response tasks should come last. It is not expected that all item types will appear in the same test. Seldom are more than a few types used, but this is the general order.

Arranging the sections of the test in this order produces a sequence that roughly approximates the complexity of the learning outcomes measured, ranging from the simple to the complex. It is then merely a matter of grouping the items within each item type. For this purpose, items that measure similar outcomes should be placed together and then

Review of Test Items Selected from Item Banks

Test items selected from workbooks, teacher guides, instructor manuals, and item banks are seldom appropriate for use without modification. Thus, before they are used in a classroom test, they should be screened and modified to fit the local instructional program. Both the checklists for reviewing test items and tasks, presented in Chapters 7 through 11, and the list of review questions in this chapter are guides for this purpose. Our aim when selecting items for classroom use should be the same as when constructing them. We want the items to be both technically sound and relevant to what has been taught during the instruction.

arranged in order of ascending difficulty. For example, the items in the multiple-choice section might be arranged in the following order: (a) knowledge of terms, (b) knowledge of specific facts, (c) knowledge of principles, and (d) application of principles. Keeping together items that measure similar outcomes is especially helpful in determining the types of learning outcomes causing students the greatest difficulty.

If, for any reason, it is not feasible to group the items by the learning outcomes measured, then it is still desirable to arrange them in order of increasing difficulty. Beginning with the easiest items and proceeding gradually to the most difficult has a motivating effect on students. Also, encountering difficult items early in the test often causes students to spend a disproportionate amount of time on such items. If the test is long, they may be forced to omit later questions that they could easily have answered.

With the items classified by item type, the sections of the test and the items within each section can be arranged in order of increasing difficulty. Some shifts in the first four item types may be warranted by the difficulty of the items used, but the interpretive exercises and essay items certainly should be last.

In constructing classroom achievement tests, there is little to be gained by grouping test items according to content. When it appears desirable to do so, such as in separating historical periods, these divisions should be kept to a minimum.

Extended-response essay questions, by their very nature, require separate administration. Performance-based tasks or oral presentations requiring extended time or access to such resources as the library, laboratory equipment, or a computer for construction of a response also obviously need to be assigned as separate units rather than as part of a classroom test.

To summarize, the most effective method for organizing items in the typical classroom test is to (a) form sections by item type, (b) group the items within each section by the learning outcomes measured, and (c) arrange both the sections and the items within sections in an ascending order of difficulty. Use subject-matter groupings only when needed for some specific purpose.

Preparing Directions for the Test or Assessment

Teachers sometimes devote considerable time and attention to the construction and assembly of test items or a challenging performance-based assessment and then dash off directions with very little thought. In fact, many teachers include no written directions with their tests, assuming either that the items are self-explanatory or that the students are

conditioned to answering the types of items used in the test. Some teachers also use oral directions, but they frequently leave much to be desired. Whether written, oral, or both, the directions should include at least the following points (see Gronlund, 2005).

1. Purpose of the test or assessment
2. Time allowed for completing the test or performing the task
3. Directions for responding
4. How to record the answers
5. What to do about guessing for selection-type test items
6. The basis for scoring open-ended or extended responses

The amount of detail for each of these points depends mainly on the students' age level, the comprehensiveness of the test or assessment, the complexity of the items or tasks, and the students' experience with the testing or assessment procedure used. Using new item types and separate answer sheets, for example, requires much more detailed directions than do familiar items requiring students merely to circle or underline the answer.

Purpose of the Test or Assessment. The purpose of the test or assessment is usually indicated when the test is announced or at the beginning of the semester when the evaluation procedures are described as a part of the general orientation to the course. Should there be any doubt whether the purpose is clear to all students, however, it could be explained again at the time of testing or assessment. This is usually done orally. The only time a statement of the purpose of the test or assessment needs to be included in the written directions is when it is to be administered to several sections taught by different teachers. Then a written statement of purpose ensures greater uniformity.

Time Allowed for Completing the Test or Performing the Task. It is helpful to tell the students how much time they will have for the whole test or performance task and how to distribute their time among the parts. When essay questions are included, it is also good to indicate approximately how much time should be allotted to each question. This enables the students to use their time most effectively and prevents less able students from spending too much time on questions that are particularly difficult for them.

Classroom tests or assessments of achievement should generally have liberal time allowances. With a few exceptions, such as measures of fluency or special computational skills, speed is not important. Our main concern is the level of achievement each student has attained. Were it not for practical considerations like the length of class periods and the pressure of other school activities, there would be no need for any time limits with most classroom achievement tests or assessment tasks.

Judging the amount of time that students will need to complete a given test or assessment task is not simple. It depends on the types of items and tasks used, the age and ability of the students, and the complexity of the learning outcomes measured. As a rough guide, the average high school student should be able to answer two true–false items, one multiple-choice item, or one short-answer item per minute of testing time. Interpretive test items take much more time; the exact amount depends on the length and complexity of the introductory materials. The time required for essay questions and other performance-based assessment tasks can vary anywhere from a few minutes each to several class periods. Also, elementary school students generally require more time per

item than high school students, and reading skill is an important determiner of the amount of time needed by a specific group. Experienced teachers familiar with the ability and work habits of a given group of students are in the best position to judge time allotments. It is better to err in the direction of allotting too much time than to deprive some of the slower students from demonstrating their maximum levels of achievement.

Directions for Responding. The directions for each section of the test should indicate the basis for selecting or supplying the answers. With true–false, matching, and multiple-choice items, this part of the directions can be relatively simple. For example, the statement, "Select the choice that best completes the statement or answers the question" might be sufficient for multiple-choice items. When interpretive exercises are used, however, more detailed directions are necessary because the basis for the response is much more complex. The directions must clearly indicate the type of interpretation expected. As stated in Chapter 9, each interpretive exercise usually requires its own directions.

It is sometimes good to include sample test items correctly marked so that students can check their understanding of the basis for answering. This practice is especially helpful to elementary school students and to students at other levels when complex item types are used.

As noted earlier, essay questions and other performance-based assessment tasks frequently require special directions concerning the type of response expected. If the selection and organization of ideas are emphasized, for example, this should be indicated to the students so that they have a more adequate basis for responding.

Procedure for Recording Answers. Answers may be recorded on the test form itself or on separate answer sheets. If the test is short, the number of students taking the test is small, or the students are relatively young, then answers are generally recorded directly on the test paper. For most other situations, separate answer sheets are preferred because they reduce the time needed for scoring, and they make it possible to use the test papers over again. The latter feature is especially useful when the test is to be given to students in different sections of the same course.

Directions for recording the answer on the test paper itself can be relatively simple. With selection items, it is merely a matter of instructing the students to circle, underline, or check the letter indicating the correct answer. For students in the primary grades, it is usually better to ask them to mark the answer directly by drawing a line under it. With supply items, the directions should indicate where to put the answer and the units in which it is to be expressed if the answer is numerical.

Separate answer sheets are easily constructed, and the directions for their use can be placed on the test paper or on the answer sheet itself. A common type of teacher-made sheet is shown in Figure 14.1. The directions on this sheet are rather general, as they must cover instructions for recording various types of answers. Students are instructed to cross out rather than circle the letters indicating the correct answers to facilitate scoring with a stencil key. Circled letters cannot be readily seen through holes in a stencil.

Special answer sheets for machine scoring can be used with classroom tests, but there is no advantage in using them unless machine scoring facilities are readily available and the number of papers to be scored warrants the expense. When machine scoring is used, special directions should be obtained from the company supplying the scoring service.

Course _____ Name _____

Section _____ Date _____

Test _____ Score: Part I _____

Part II _____

Total _____

Directions: Read all directions on the test paper carefully and follow them exactly. For each test item, indicate your answer on this sheet by crossing out the appropriate letter (X) or filling in the appropriate blank. Be sure that the number on the answer sheet is the same as the number of the test item your are answering.

True–False		Multiple-Choice		Short Answer	
Item	*Answer*	*Item*	*Answer*	*Item*	*Answer*
1	T F	21	A B C D E	41	_____
2	T F	22	A B C D E	42	_____
3	T F	23	A B C D E	43	_____
4	T F	24	A B C D E	44	_____
..	.. ..	..	. . . . .	..	.. ..

Figure 14.1
Top portion of a teacher-made answer sheet

What to Do About Guessing for Selection-Type Items. When selection-type items are used, the directions should tell students what to do when they are uncertain of the answer. Should they guess or omit the item? If no instructions are given on this point, the bold students will guess freely, whereas others will answer only those items of which they are fairly certain. The bold students will select some correct answers just by lucky guesses, and thus their scores will be higher than they should be. On the other hand, if the students are instructed "Do not guess" or "Answer only those items of which you are certain," the more timid students will omit many items they could answer correctly. Such students are not very certain about anything, which prevents them from responding even when they are reasonably sure of the answers. With these directions, the bold students will continue to guess, although possibly not quite so wildly.

As Cronbach (1990) pointed out, the tendency to guess or not to guess when in doubt about an item is determined by personality factors and cannot be entirely eliminated by directions that caution against guessing or that promise penalties to those who do guess. The only way to eliminate variations in the tendency to guess is to instruct students to answer every item. When this is done, no student is given a special advantage, and it is unnecessary to correct for guessing in the scoring. Directions such as the following are usually sufficient to communicate this to the students: "Because your score is the number right, be sure to answer every item."

Some teachers object to such directions on the grounds that encouraging guessing is undesirable from an educational standpoint. Most responses to doubtful items are not wild guesses, however, but are guided by some information and understanding. In this respect,

they are not too different from the informed guesses we make when we predict weather, judge the possible consequences of a decision, or choose one course of action over another. Problem solving always involves a certain amount of this type of informed guessing.

A more defensible objection to directions that encourage guessing is that the chance errors introduced into the test scores lower the accuracy of measurement. Although this is certainly objectionable, it probably has less influence on the validity of the results than does the systematic advantage given to the bold guessers by the "do not guess" directions.

For liberally timed classroom tests, the "answer every item" directions are favored. For speed tests and when teachers want to discourage guessing, however, directions such as the following are a good compromise: "Answer all items for which you can find some reasonable basis for answering, even though you are not completely sure of the answer. Do not guess wildly, though, because there will be a correction for guessing."

The Basis for Scoring Open-Ended or Extended Responses. For tasks requiring extended or open-ended written responses, it is important to tell students the basis for scoring. If there are several essay questions, for example, the number of points possible for the response to each question should be indicated. The importance attached to factors such as factual accuracy, organization, comprehensiveness, persuasiveness, and originality can be indicated. Students should also be informed if their responses will be graded for mechanics (see the "Guidelines" box).

GUIDELINES

Helping Students Prepare for Tests Assessments

General Preparation

1. Suggest ways of studying.
2. Give practice tasks like those to be used.
3. Teach test-taking skills.
4. Teach how to write well-organized essay answers.
5. Stress the value of tests and assessments for improving learning.

Preparation for Each Test or Assessment

1. Announce in advance when the test or assessment will be given.
2. Describe the conditions of administration (e.g., 1-hour closed book).
3. Describe the length and the types of items or tasks to be used (20 multiple-choice, three essay items, or one extended-response performance task).
4. Describe the content and type of performance to be covered (a table of specifications is useful for this).
5. Describe how the test or assessment will be scored and how the results will be used.
6. Give the students sample items and tasks similar to those to be used (use a short practice test or present items orally and discuss responses).
7. Relieve anxiety by using a positive approach in describing the test or assessment and its usefulness.

Reproducing the Test

In preparing the test materials for reproduction, it is important that the items be spaced and arranged so that they can be read, answered, and scored with the least amount of difficulty. Cramming too many test items onto a page is poor economy. What little paper is saved will not make up for the time and confusion that results during the administration and scoring of the test.

All test items should have generous borders. Multiple-choice items should have the alternatives listed in a vertical column beneath the stem of the item rather than across the page. Items should not be split, with parts of the item on two different pages. With interpretive exercises, the introductory materials can sometimes be placed on a facing page or separate sheet, with all the items referring to it on a single page.

Unless a separate answer sheet is used, the space for answering should be down one side of the page, preferably the left. The most convenient method of response is circling the letter of the correct answer. With this arrangement, scoring is simply a matter of placing a strip scoring key beside the column of answers.

Test items should be numbered consecutively throughout the test. Each test item will need to be identified during discussion of the test and for other purposes, such as item analysis. When separate answer sheets are used, consecutive numbering is, of course, indispensable.

It is desirable to proofread the entire test or assessment before it is administered. Charts, graphs, and other pictorial material must be checked to ensure that the reproduction has been accurate and the details are clear.

ADMINISTERING AND SCORING CLASSROOM TESTS AND ASSESSMENTS

The same care that went into the preparation of the test or assessment should be carried over into its administration and scoring. Here we are concerned with (a) providing optimum conditions for obtaining the students' responses, and (b) selecting convenient and accurate procedures for scoring the results.

Administration

The guiding principle in administering any classroom test or assessment is that all students must be given a fair chance to demonstrate their achievement of the learning outcomes being measured. This means a physical and psychological environment conducive to their best efforts and the control of factors that might interfere with valid measurement.

Physical conditions such as adequate work space, quiet, proper light and ventilation, and comfortable temperature are sufficiently familiar to teachers to warrant little attention here. Of greater importance but frequently neglected are the psychological conditions influencing results. Students will not perform at their best if they are tense and anxious during testing. The following may create excessive test anxiety.

1. Threatening students with tests if they do not behave
2. Warning students to do their best "because this test is important"
3. Telling students they must work fast in order to finish on time
4. Threatening dire consequences if they fail

The antidote to test anxiety is to convey to the students, by both word and deed, that the test and assessment results are to be used to help them improve their learning. They also should be reassured that the time limits are adequate to allow them to complete the test or assessment tasks. This, of course, assumes that the test and assessment results will be used to improve learning and that the time limits are adequate.

The time of testing can also influence the results. If tests are administered just before the "big game" or the "big dance," the results may not be representative. Furthermore, for some students, fatigue, the onset of illness, or worry about a particular problem may prevent maximum performance. Arranging the time of testing accordingly and permitting its postponement when appropriate can enhance the validity of results.

Actual administration is relatively simple because a properly prepared test or assessment is practically self-administering. Oral directions, if used, should be presented clearly. Any sample problems or illustrations put on the board should be kept brief and simple. Beyond this, suggestions for administration consist mainly of things to avoid.

1. **Do not talk unnecessarily before letting students start working.** When a teacher announces that there will be "a full 40 minutes" to complete the test and then talks for the first 10 minutes, students feel that they are being unfairly deprived of testing time. Besides, just before a test is no time to make assignments, admonish the class, or introduce next week's topic. Students are mentally set for the test and will ignore anything not pertaining to the test for fear it will hinder their recall of information needed to answer the questions. Thus, the well-intentioned remarks merely increase anxiety toward the test and create hostility toward the teacher.

2. **Keep interruptions to a minimum.** At times, a student will ask to have an ambiguous item clarified, and it may be beneficial to explain the item to the entire group at the same time. Such interruptions are necessary but should be kept to a minimum. All other distractions outside and inside the classroom should, of course, also be eliminated when possible. It is sometimes helpful to hang a "Do not disturb—TESTING" sign outside the door.

3. **Avoid giving hints to students who ask about individual items.** If the item is ambiguous, it should be clarified for the entire group, as indicated earlier. If it is not ambiguous, refrain from helping the student answer it. Refraining from giving hints to students who ask for help is especially difficult for beginning teachers; but giving unfair aid to some students (the bold, the apple polishers, and so on) decreases the validity of the results and lowers class morale.

4. **Discourage cheating, if necessary.** When there is good teacher–student rapport and the students view tests as helpful rather than harmful, cheating is usually not a problem. Under other conditions, however, it might be necessary to discourage cheating by special seating arrangements and careful supervision. Receiving unauthorized help from other students during a test has the same deleterious effect on validity and class morale as does receiving special hints from the teacher. We are interested in students doing their best; but

GUIDELINES

Steps to Prevent Cheating

1. Take special precautions to keep the test secure during preparation, storage, and administration.
2. Have students clear off the tops of their desks (for adequate work space and to prevent use of notes).
3. If scratch paper is used (e.g., for math problems), have it turned in with the test.
4. Proctor the testing session carefully (e.g., walk around the room periodically and observe how the students are doing).
5. Use special seating arrangements, if possible (e.g., leave an empty row of seats between students).
6. Use two forms of the test and give a different form to each row of students (for this purpose, use the same test but simply rearrange the order of the items for the second form).
7. Prepare tests that students will view as relevant, fair, and useful.
8. Create and maintain a positive attitude concerning the value of tests for improving learning.

for valid results, their scores must be based on their own unaided efforts. See the "Guidelines" box.

Scoring the Test

Procedures for scoring performance-based assessments were described in Chapter 11. Here we discuss scoring objective items.

If the students' answers are recorded on the test paper itself, a scoring key can be made by marking the correct answers on a blank copy of the test. Scoring then is simply a matter of comparing the columns of answers on this master copy with the columns of answers on each student's paper. A strip key, which consists merely of strips of paper on which the columns of answers are recorded, may also be used if more convenient. These can easily be prepared by cutting the columns of answers from the master copy of the test and mounting them on strips of cardboard cut from manila folders.

When separate answer sheets are used, a scoring stencil is most convenient. This is a blank answer sheet with holes punched where the correct answers should appear. The stencil is laid over each answer sheet, and the number of marks appearing through the holes are counted. When this type of scoring procedure is used, each test paper should also be scanned to make certain that only one answer was marked for each item. Any item containing more than one answer should be eliminated from the scoring.

As each test paper is scored, mark each item that is answered incorrectly. With multiple-choice items, a good practice is to draw a red line through the correct answer of the missed items rather than through the student's wrong answers. This will indicate to the student those items missed and at the same time will indicate the correct answers. Time

will be saved and confusion avoided during discussion of the test. Marking the correct answers of missed items is especially simple with a scoring stencil. When no mark appears through a hole in the stencil, a red line is drawn across the hole.

In scoring objective tests, each correct answer is usually counted as 1 point because an arbitrary weighing of items makes little difference in the students' final scores. If some items are counted as 2 points, some 1 point, and some 0.5 point, the scoring will be more complicated without any accompanying benefits. Scores based on such weightings will be similar to the simpler procedure of counting each item as 1 point. When a test consists of a combination of objective items and a few more time-consuming essay questions, however, more than a single point is needed to distinguish several levels of response and to reflect the disproportionate time devoted to each of the essay questions.

When students are told to answer every item on the test, a student's score is simply the number of items answered correctly. There is no need to consider wrong answers or to correct for guessing. When all students answer every item on a test, the rank of the students' scores will be the same whether the number right or a correction for guessing is used.

See the box "Correction for Guessing" for a simple formula that is sometimes used. The formula is based on the questionable assumption that students either know the answer or guess it at random. It is not needed when students are allowed sufficient time to respond to all the items on the test. Thus, it is recommended that it *not* be used with the ordinary classroom test. The only exception is when the test is speeded to the extent that students complete different numbers of items. Here its use is defensible because students can increase their scores appreciably by rapidly (and blindly) guessing at the remaining untried items just before the testing period ends.

Correction for Guessing

Correcting for guessing is usually done when students do not have sufficient time to complete all items on the test and when they have been instructed that there will be a penalty for guessing. The most common formula used for this purpose is the following:

$$\text{Score} = \text{Right} - \text{Wrong}/(n - 1)$$

In this formula, n is the number of alternatives for an item. Thus, the formula applies to various selection-type items as follows:

True–False Items

$$\text{Score} = \text{Right} - \text{Wrong}/(2 - 1)$$

(or)

$$\text{Score} = \text{Right} - \text{Wrong}$$

Multiple-Choice Items

$$\text{Score} = \text{Right} - \text{Wrong}$$

Three alternatives	Score = Right – Wrong/2
Four alternatives	Score = Right – Wrong/3
Five alternatives	Score = Right – Wrong/4

APPRAISING CLASSROOM TESTS AND ASSESSMENTS

Before a classroom test or assessment has been administered, it should be evaluated according to the points discussed earlier. The most important of these points are listed in the checklist "Evaluating the Classroom Assessment." Most of these questions also apply to a performance-based assessment. A *yes* response to each of these questions indicates that the test or the assessment has been carefully prepared and will probably function effectively.

After a test or assessment has been scored and the students have discussed the results, it is often simply discarded. Except for the students' criticism during class discussion, which helps identify some of the defective items or ambiguities in a task, the teacher has little evidence concerning the quality of the test or assessment that was used. Much of the careful planning and hard work that went into the preparation of the test or assessment is wasted. A better procedure is to appraise the effectiveness of the test items and assessment tasks and to build a file of high-quality items and tasks for future use.

Determining Item and Task Effectiveness

The effectiveness of each test item can be determined by analyzing student responses to it. Item analysis is generally associated with a norm-referenced perspective. This is natural because the results of an item analysis can be used to select items of desired difficulty that best discriminate between high- and low-achieving students. Selection on these grounds is not relevant from a criterion-referenced perspective. From both perspectives, however, the results of an item analysis can be useful in identifying faulty items and can provide information about student misconceptions and topics that need additional work.

Item analysis is usually designed to answer questions such as the following:

1. Did the item function as intended?
2. Was the test item of appropriate difficulty?
3. Was the test item free of irrelevant clues and other defects?
4. Were the distracters effective (in multiple-choice items)?

Answers to all but the second question are relevant in constructing future tests based on either a norm-referenced or a criterion-referenced perspective. The answer to the second question is relevant only when planning future norm-referenced tests; however, it is relevant in instructional planning regardless of perspective.

Answers to such questions are of obvious value in selecting or revising items for future use. The benefits of item analysis are not limited to the improvement of individual test items, however. There are a number of fringe benefits of special value to classroom teachers. The most important of these are the following:

1. **Item-analysis data provide a basis for efficient class discussion of the test results.** Knowing how effectively each item or task functioned in measuring achievement makes it possible to confine the discussion to those areas most helpful to students. Misinformation and misunderstandings reflected in the choice of particular distracters on multiple-choice problems or frequently repeated errors on performance tasks can be corrected, thereby enhancing the instructional value of the assessment. Item analysis will also expose

CHECKLIST

Evaluating the Classroom Assessment

Adequacy of Assessment Plan

	Yes	No
1. Does the assessment plan adequately describe the instructional objectives and the content to be measured?	____	____
2. Does the assessment plan clearly indicate the relative emphasis to be given to each objective and each content area?	____	____

Adequacy of Test Items and Assessment Tasks

	Yes	No
3. Is the format of each item and task suitable for the learning outcome being measured *(appropriateness)*?	____	____
4. Does each item or task require pupils to demonstrate the performance described in the specific learning outcome it measures *(relevance)*?	____	____
5. Does each item or task present a clear and definite task to be performed *(clarity)*?	____	____
6. Is each item or task presented in simple, readable language and free from excessive verbiage *(conciseness)*?	____	____
7. Does each item or task provide an appropriate challenge *(ideal difficulty)*?	____	____
8. Does each item or task have an answer that would be agreed upon by experts *(correctness)*?	____	____
9. Is there a clear basis for awarding partial credit on items or tasks with multiple points *(scoring rubric)*?	____	____
10. Is each item or task free from technical errors and irrelevant clues *(technical soundness)*?	____	____
11. Is each test item free from racial, ethnic, and gender bias *(cultural fairness)*?	____	____
12. Is each test item independent of the other items in the test *(independence)*?	____	____
13. Is there an adequate number of test items for each learning outcome *(sample adequacy)*?	____	____

Adequacy of Test Format and Directions

	Yes	No
14. Are test items of the same type grouped together in the test (or within sections of the test)?	____	____

	Yes	No
15. Are the test items arranged from easy to more difficult within sections of the test and the test as a whole?	___	___
16. Are the test items numbered in sequence?	___	___
17. Is the answer space clearly indicated (on the test itself or on a separate answer sheet), and is each answer space related to its corresponding test item?	___	___
18. Are the correct answers distributed in such a way that there is no detectable pattern?	___	___
19. Is the test material well spaced, legible, and free of typographical errors?	___	___
20. Are there directions for each section of the test and the test as a whole?	___	___
21. Are the directions clear and concise?	___	___

technical defects in items and tasks. It can also suggest needed changes in scoring rubrics. During discussion, defective items can be pointed out to students, saving much time and heated discussion concerning the unfairness of these items. If an item is ambiguous and two answers can be defended equally well, both answers should be counted correct and the scoring adjusted accordingly.

2. **Item-analysis data provide a basis for remedial work.** Although discussing the test results in class can clarify and correct many specific points, item analysis frequently brings to light general areas of weakness requiring more extended attention. It is often informative to compare actual student performance on a task to the performance expected based on the teacher's notion of how challenging the task would be for students. Performance that is much worse than expected may suggest the need to revisit particular critical concepts or topics. In a mathematics test, for example, item analysis may reveal that the students are fairly proficient in mathematics skills but are having difficulty with problems requiring the application of these skills. In other subjects, item analysis may indicate a general weakness in knowledge of technical vocabulary, in an understanding of principles, or in the ability to interpret data. Such information makes it possible to focus remedial work directly on the particular areas of weakness.

3. **Item-analysis data provide a basis for the general improvement of classroom instruction.** In addition to the preceding uses, item-analysis data can assist in evaluating appropriateness of the learning outcomes and course content for the particular students being taught. For example, material that is consistently too simple or too difficult might suggest curriculum revisions or shifts in teaching emphasis. Similarly, errors in student thinking that persistently appear in item-analysis data might direct attention to the need for more effective teaching procedures. In these and similar ways, item-analysis data can reveal instructional weaknesses and clues for improvement.

4. **Item-analysis procedures provide a basis for increased skill in test construction.** Item analysis reveals ambiguities, clues, ineffective distracters, and other technical defects that

were missed during the test's preparation. This information is used directly in revising the test items for future use. In addition to the improvement of the specific items, however, we derive benefits from the procedure itself. As we analyze students' responses to items, we become increasingly aware of technical defects and what causes them. When revising the items, we gain experience in rewording statements so that they are clear, rewriting distracters so that they are more plausible, and modifying items so that they are at a more appropriate level of difficulty. As a consequence, our general test construction skills improve.

Simplified Item-Analysis Procedures

A simplified form of item analysis is all that is necessary or warranted for classroom tests. Because most classroom groups consist of 20 to 40 students, an especially useful procedure is to compare the responses of the 10 highest-scoring students with the responses of the 10 lowest-scoring students. As we will see later, keeping the upper and lower groups at 10 students each simplifies the interpretation of the results. It also is a reasonable number for analysis in groups of 20 to 40 students. For example, with a small classroom group, like that of 20 or fewer students, it is best to use the upper and lower halves to obtain dependable data, whereas with a larger group, like that of 40 students, use of the upper and lower 25% is quite satisfactory. For more refined analysis, the upper and lower 27% is often recommended, and most statistical guides are based on that percentage.

To illustrate the method of item analysis, suppose that we have just finished scoring 32 test papers for a sixth-grade science unit on weather. Our item analysis might then proceed as follows:

1. Rank the 32 test papers in order from the highest to the lowest score.
2. Select the 10 papers within the highest total scores and the 10 papers with the lowest total scores.
3. Put aside the middle 12 papers, as they will not be used in the analysis.
4. For each test item, tabulate the number of students in the upper and lower groups who selected each alternative. This tabulation can be made directly on the test paper or on the test item record, as shown in Figure 14.2.
5. Compute the difficulty of each item (percentage of students who got the item right).
6. Compute the discriminating power of each item (difference between the number of students in the upper and lower groups who got the item right).
7. Evaluate the effectiveness of distracters in each item (attractiveness of the incorrect alternatives).

The first steps of this procedure are merely a convenient tabulation of student responses from which we can readily determine item difficulty, item discriminating power, and the effectiveness of each distracter. This latter information can frequently be obtained simply by inspecting the item-analysis data. Note that in Figure 14.2, for example, when the item was used in the spring of 2006, eight students in the upper group and four students in the lower group selected the correct alternative, B. Thus, 12 of the 20 students

COURSE Science UNIT Weather

OBJECTIVE Identifies use of instruments

ITEM

Which of the following is most useful in weather forecasting?

A. Anemometer
*B. Barometer
C. Thermometer
D. Rain gauge

Item-Analysis Data

Frequencies								Indices	
Dates Used	*Pupils*	*Alternatives* A	(B)	C	D	E	*Omits*	*Difficulty*	*Discrimination*
4/25/06	Upper 10	1	8	0	1		0		
	Lower 10	2	4	1	3		0	60%	.40
4/30/08	Upper 10	0	10	0	0		0		
	Lower 10	1	3	2	2		1	65%	.70
	Upper 10								
	Lower 10								
	Upper 10								
	Lower 10								

Comment:

Figure 14.2
Test item record with item-analysis data

(difficulty = 60%) got the item right, indicating that the item has a moderate difficulty. When the same item was used in 2008, the difficulty was similar (65%) based on 10 students in the upper group and 3 in the lower group for a total of 13 of 20 who selected the correct option (B).

Because more students in the upper group than in the lower group got the item right, it is discriminating positively in both years. That is, it is distinguishing between high and low achievers (as determined by the total test score). The .40 for the 2006 administration equals the difference in the proportion of students in the upper and lower groups who answered the item correctly (8/10 – 4/10 or .8 –.4). In 2008, the discrimination of the items was higher than in 2006 (.70 = 1.00 – .30, the proportions in the upper and lower groups answering the item correctly in 2008). Finally, because all the alternatives were selected by some of the students in the lower group, the distracters (alternatives A, C, and D) appear to be operating effectively.

From a norm-referenced perspective, the fact that in 2008 all 10 students with the highest test scores but only 3 of the 10 with the lowest scores answered the item correctly makes this item a good candidate for use in the future because it helps discriminate between high- and low-achieving students. Although this would not be a basis for selecting an item for a criterion-referenced test, it does provide an indication that the item is keyed correctly and that it is not being misinterpreted by the higher-achieving students. It also provides an indication that some students do not understand the uses of instruments listed.

Although item analysis by inspection will reveal the general effectiveness of a test item and is satisfactory for most classroom purposes, it is sometimes useful to obtain a more precise estimate of item difficulty and discriminating power. This can be done by applying relatively simple formulas to the item-analysis data.

Computing Item Difficulty. The difficulty of a test item that is scored right or wrong is indicated by the percentage of students who get the item right. Hence, we can compute item difficulty (P) by means of the following formula, in which R equals the number of students who got the item right and T equals the total number of students who tried the item:

$$P = 100R/T$$

Applying this formula to the item-analysis data in Figure 14.2, our index of item difficulty is 65% for the April 30, 2008, test as follows:

$$P = 100 * 13/20 = 65\%$$

In computing item difficulty from item-analysis data, our calculation is based on the upper and lower groups only. We assume that the responses of students in the middle group follow essentially the same pattern. This estimate of difficulty is sufficiently accurate for classroom use and is easily obtained because the needed figures can be taken directly from the item-analysis data.

Note that because our item analysis is based on 10 in the upper group and 10 in the lower group, all we need to do to obtain item difficulty is to divide the number getting it right by 2 (13/2 = 6.5), move the decimal point one place to the right (65), and add the percent sign (65%). In other words, 13 of 20 is the same as 6.5 of 10, which is 65%. In April 2006, when 12 students got the item right, item difficulty was 6 of 10 (12/2 = 6), or 60%. This may seem a bit confusing at first, but once you grasp the idea, you can compute item difficulty very quickly. As noted earlier, the ease of interpreting item statistics is one of the advantages of using 10 in each group. If more (or fewer) than 10 are used, the formula for computing item difficulty is the same, but it is much more difficult to compute the results mentally.

Similar calculations are used for a task scored 0, 1, 2, or 3 to get the overall mean for the students in the upper and lower groups combined. That is, the average would simply be the sum of the means for the upper and lower groups divided by 2.

Computing Item Discriminating Power. As we have already stated, an item discriminates positively if more students in the upper group than the lower group get the item right. Positive discrimination indicates that the item is discriminating in the

same direction as the total test score. Because we assume that the total test score reflects achievement of desired objectives, we would like all our test items to show positive discrimination.

The discriminating power of an achievement test item refers to the degree to which it discriminates between students with high and low achievement. Item-discriminating power *(D)* can be obtained by subtracting the number of students in the lower group who get the item right *(RL)* from the number of students in the upper group who get the item right *(RU)* and dividing by one half the total number of students included in the item analysis *(.5T)*. Summarized in formula form, it is as follows:

$$D = (RU - RL)/(.5T)$$

Applying this formula to the item-analysis data for April 2008 in Figure 14.2, we obtain an index of discriminating power of .70 as follows:

$$D = (10 - 3)/10 = .70$$

This indicates approximately average discriminating power. An item with maximum positive discriminating power is one in which all students in the upper group get the item right and all the students in the lower group get the item wrong. This results in an index of 1.00, as follows:

$$D = (10 - 0)/10 = 1.00$$

An item with no discriminating power is one in which an equal number of students in both the upper and the lower groups get the item right. This results in an index of .00, as follows:

$$D = (10 - 10)/10 = .00$$

When our item analysis is based on 10 in the upper group and 10 in the lower group, the index of discriminating power, like item difficulty, can be computed easily and quickly. All we need to do is subtract the number in the lower group who get it right from the number in the upper group who get it right (10 – 3 = 7), move the decimal point one place to the left (.7), and add a zero after it (.70). With 10 in each group, the index of discrimination is essentially the difference between the number getting it right in the two groups with the decimal point moved one place to the left. The zero is added simply because the index of discrimination is usually carried to two decimal places. With more than 10 in each group, we could not make these simple mental calculations but would have to resort to use of the formula.

Evaluating the Effectiveness of Distracters. How well each distracter is operating can be determined by inspection, so there is no need to calculate an index of effectiveness, although the formula for discriminating power can be used for this purpose. In general, a good distracter attracts more students from the lower group than the upper group. Thus, it should discriminate between the upper and lower groups in a manner opposite to that of the correct alternative. An examination of the following item-analysis data will illustrate the ease with which the effectiveness of distracters can be determined by inspection. Alternative A is the correct answer:

Alternatives	(A)	B	C	D	Omits
Upper 10	5	4	0	1	0
Lower 10	3	2	0	5	0

First, note that the item discriminates positively because five in the upper group and three in the lower group got the item right. The index of discriminating power is fairly low, however (D = .20), and this may be partly due to the ineffectiveness of some of the distracters. Alternative B is a poor distracter because it attracts more students from the upper group than from the lower group. This is most likely due to some ambiguity in the statement of the item. Alternative C is evidently not a plausible distracter because it attracted no one. Alternative D is functioning as intended, for it attracts a larger proportion of students from the lower group. Thus, the discriminating power of this item can probably be improved by removing any ambiguity in the statement of the item and revising or replacing alternatives B and C. The specific changes must, of course, be based on an inspection of the test item itself; item-analysis data merely indicate poorly functioning items, not the cause of the poor functioning.

In some cases, an examination of the test item will reveal no obvious error in the structure of the item and it may be best to try it with a second group. The number of cases involved is so small that considerable variation in student response can be expected from one group to another. A casual comment by the teacher or some other classroom event may cause students to select or reject a particular alternative.

Recording Item-Analysis Data on the Test Paper. There is some advantage in recording item-analysis data directly on the test paper that was used as a scoring key, as shown in Figure 14.3, and making the calculations mentally. These mental calculations for the two items in Figure 14.3 are summarized in Figure 14.4. Thus, during discussion of the test results, you can quickly judge the difficulty and discriminating power of each item and the effectiveness of the distracters. This will help determine how much discussion to devote to any particular item, the types of misconceptions students may have (by the distracters selected), and which items are so defective that they might be discounted. See "Item Analysis by Computer" for a look at another way to analyze test items.

Cautions in Interpreting Item-Analysis Results

Item analysis is a quick, simple technique for appraising the effectiveness of individual test items. The information from such an analysis is limited in many ways, however, and must be interpreted accordingly. Observe the following major cautions.

1. **Item discriminating power does not indicate item validity.** In our description of item analysis, we used the total test score as a basis for selecting the upper group (high achievers) and the lower group (low achievers). This is the most common procedure because comparable measures of achievement are usually not available. Ideally, we would examine each test item in relation to some independent measure of achievement. However, the best measure of the particular achievement we are interested in assessing is usually the total score on the achievement test we have constructed because each classroom test

WEATHER UNIT

Name _____ Date _____

Directions. This test will measure what you have learned during the unit on weather. There are 40 objective questions in the test. You will have the entire class period to complete it.

For each question there are several possible answers. Select the *best* answer and indicate it by encircling the letter of your answer.

Your score will be the number of questions answered correctly so *be sure to answer every question*.

KNOWLEDGE OF FACTS

U	L	
		1. Which of these instruments is used to measure humidity?
0	1	A. Anemometer
0	1	B. Barometer
10	8	(C.) Hygrometer
0	0	D. Thermometer
U	L	2. What does the Beaufort scale indicate on a weather map?
1	2	A. Air pressure
0	1	B. Air temperature
0	1	C. Precipitation
9	6	(D.) Wind velocity

Figure 14.3
Sample test scoring key with item-analysis data added (U = Upper 10 pupils, L = Lower 10 pupils)

Item Difficulty Index

Steps (using numbers to left of answer)	*Item 1*	*Item 2*
1. Add $U + L$ and divide by 2.	18/2 = 9	15/2 = 7.5
2. Move decimal point one to the right.	90	75
3. Add the percent sign.	90%	75%

Item Discrimination Index

Steps (using numbers to left of answer)	*Item 1*	*Item 2*
1. Subtract $U - L$.	10 − 8 = 2	9 − 6 = 3
2. Move decimal point one place to the left.	.2	.3
3. Add a zero after the number.	.20	.30

Figure 14.4
Illustrative item-analysis calculations from data in Figure 14.3

is related to specific instructional objectives and course content. Even standardized tests in the same content area are usually inadequate as independent criteria because they are aimed at more general objectives than those measured by a classroom test in a particular course.

Using the total score from our classroom test as a basis for selecting high and low achievers is perfectly legitimate as long as we remember that we are using an internal criterion. In doing so, our item analysis offers evidence concerning the internal consistency of the test rather than its validity. That is, we are determining how effectively each test item is measuring whatever the whole test is measuring. Such item-analysis data can be interpreted as evidence of item validity only when the validity of the total test has been proven or can be legitimately assumed. This is seldom possible with classroom tests, so we must be satisfied with a more limited interpretation of our item-analysis data.

2. A low index of discriminating power does not necessarily indicate a defective item. Items that discriminate poorly between high and low achievers should be examined for the possible presence of ambiguity, clues, and other technical defects. If none is found and the items measure an important learning outcome, they should be retained for future use. Any item that discriminates in a positive direction can contribute to the measurement of student achievement, and low indexes of discrimination are frequently obtained for reasons other than technical defects.

Classroom achievement tests are usually designed to measure several different types of learning outcomes (knowledge, understanding, application, and so on). When this is the case, test items that represent an area receiving relatively little emphasis will tend to have poor discriminating power. For example, if a test has 40 items measuring knowledge of facts and 10 items measuring understanding, the latter items can be expected to have low indexes of discrimination, because the items measuring understanding have less representation in the total test score and there is typically a low correlation between measures of knowledge and measures of understanding. Low indexes of discrimination here merely indicate that these items are measuring something different from what the major part of the test is measuring. Removing such items from the test would make it a more homogeneous measure of knowledge outcomes, but it would also damage the test's validity because it would no longer measure learning outcomes in the understanding area. Because most classroom tests measure a variety of types of learning outcomes, low positive indexes of discrimination are the rule rather than the exception.

Another factor that influences discriminating power is the difficulty of the item. Those items at the 50% level of difficulty make maximum discriminating power possible because only at this level of difficulty can all students in the upper half of the group get the item right and all students in the lower half get it wrong. The 50% level of difficulty does not guarantee maximum discriminating power but merely makes it possible. If half the students in the upper group and half the students in the lower group got the item right, the level of difficulty would still be 50%, but the index of discrimination would be zero. As we move away from the 50% level of difficulty toward easier or more difficult items, the index of discriminating power becomes smaller. Thus, items that are very easy or very difficult have low indexes of discriminating power. Sometimes it is necessary or desirable to retain such items, however, in order to measure a representative sample of learning out-

comes and course content. To summarize, a low index of discriminating power should alert us to the possible presence of technical defects in a test item but should not cause us to discard an otherwise worthwhile item. A well-constructed achievement test will, of necessity, contain items with low discriminating power; to discard them would result in a less, rather than more, valid test.

3. **Item-analysis data from small samples are highly tentative.** Item-analysis procedures focus our attention so directly on a test item's difficulty and discriminating power that we are commonly misled into believing that these are fixed, unchanging characteristics. This, of course, is not true. Item-analysis data will vary from one group to another, depending

Item Analysis by Computer

Many schools now have computers (or have access to them) that can both score and analyze tests. The computer printout will provide item-analysis information, a reliability coefficient, standard error of measurement for the test, and various other types of information concerning the performance of the individuals tested and the characteristics of the test. The nature of the information depends on the sophistication of the computer and the program that is used.

When item analysis is done by computer, the scores of the entire group are usually used rather than just the scores of the upper and lower groups. The total set of scores might be divided into two, three, four, or five levels, depending on the size of the group and the types of analyses. Item-analysis data on a computer printout based on 50 pupils might appear as follows, for each item.

Item-Response Pattern

Item 1	A	B	(C)	D	E	Omit	Total	Item Statistics
Upper 30%	1	1	12	1	0	0	15	Difficulty 60%
Middle 40%	2	2	12	3	1	0	20	Discrimination .40
Lower 30%	2	3	6	3	1	0	15	
Total	5	6	30	7	2	0	50	

The item-response data indicate how many pupils, at each level, selected the correct answer (C) and how many selected each of the distracters.The item statistics at the right indicate the index of difficulty and the index of discrimination for this item. Some computer programs report only the item statistics, but the item-response pattern is especially valuable for evaluating the effectiveness of the distracters and planning for item revision. Alternative E, for example, should be examined to determine whether it can be replaced by a more effective distracter because it is rarely selected.

The following Web sites provide information about two of the many item-analysis computer programs that are available.

ITEMAN at
http://www.assess.com
Remark Products at
http://www.gravie.com/remark/

on the students' level of ability, educational background, and type of instruction they have had. Add to this the small number of students available for analyzing the items in our classroom tests, and the tentative nature of our item-analysis data becomes readily apparent. If just a few students change their responses, our indexes of difficulty and discriminating power can be increased or decreased by a considerable amount.

The tentative nature of item-analysis data should discourage us from making fine distinctions among items on the basis of indexes of difficulty and discriminating power. If an item is discriminating in a positive direction, all the alternatives are functioning effectively, and it has no apparent defects, then it can be considered satisfactory from a technical standpoint. The important question then is **not** how high the index of discriminating power is but whether the item measures an important learning outcome. In the final analysis, the worth of an achievement test item must be based on logical rather than statistical considerations.

When used with norm-referenced classroom tests, item analysis provides us with a general appraisal of the functional effectiveness of the test items, a means for detecting defects, and a method for identifying instructional weaknesses. For these purposes, the tentative nature of item-analysis data is relatively unimportant. When we record indexes of item difficulty or discriminating power on item records for future use, we should interpret them as rough approximations only. As such, they are still superior to our unaided estimates of item difficulty and discriminating power.

Application of Item-Analysis Principles with Performance-Based Assessments

Item-analysis procedures have somewhat limited applicability with performance-based assessments, primarily because such assessments generally contain a relatively small number of tasks. If the assessment has several tasks, however, the general principles can be readily adapted for use. One necessary modification results from the fact that scores on performance-based tasks almost always involve more than a simple 0 or 1. For example, each task might have possible scores of 0, 1, 2, 3, or 4. Still, a comparison of the individual task scores for the 10 highest and 10 lowest scoring students can be useful.

Suppose, for example, that a performance assessment consisted of five separate tasks, each of which had possible scores ranging from a low of 0 for no response or a response that was unrelated to the task, to a 4 for a complete and well-elaborated response. Possible total scores for the set of five tasks would range from 0 to 20. As before, the total scores would be ranked to identify the 10 highest and the 10 lowest scores. The individual task scores for these two groups of students would then be summarized as illustrated in Figure 14.5. The 10 highest-achieving students generally have higher scores on each item than the 10 lowest-achieving students. The higher average score on Task 1 for the upper 10 students than for the lower 10 indicates that the task discriminates between these two groups. The equal means on Task 2 for the two groups of students indicate that the latter task does not discriminate.

A couple of possible reasons for results such as those shown in Figure 14.5 need to be considered. It is possible that Task 2 simply has less similarity than Task 1 to the remaining tasks; that is, it calls for different skills and abilities than the other four tasks. It

Task 1

Score	0	1	2	3	4	Average
Upper 10	0	0	1	4	5	3.4
Lower 10	1	3	4	2	0	1.7

Task 2

Score	0	1	2	3	4	Average
Upper 10	2	2	3	1	1	1.5
Lower 10	3	2	2	3	0	1.5

Figure 14.5
Illustrative analysis of scores on two tasks

may be that the type of performance expected on Task 2 is ambiguous. If careful review of the task in comparison to the other four tasks leads to the first conclusion, then there is no need to revise or discard the task. If a review leads to the second conclusion, however, then the task would need to be either discarded or revised to clarify what is intended.

BUILDING A FILE OF EFFECTIVE ITEMS AND TASKS

A file of effective items and tasks can be built and maintained easily if items and tasks are recorded on records like the one shown in Figure 14.2. By indicating on the record both the objective and the content area being measured, it is possible to file the records under both headings. Course content can be used as major categories, with the objectives forming the subcategories. For example, the item in Figure 14.2 measures knowledge of weather instruments, so it is placed in the first category under weather instruments as follows:

Weather Instruments

Knowledge

Understanding

Application

This type of filing system makes it possible to select items or tasks in accordance with any table of specifications in the particular area covered by the file. See the box "Item Banking by Computer" for a description of another way to build an assessment bank.

Building a file of effective items and tasks is a little like building a bank account. The first several years are concerned mainly with making deposits; withdrawals must be delayed until a sufficient reserve is accumulated. Thus, items and tasks are recorded on records as they are constructed; information from analyses of student responses is added after the items and tasks have been used, and then the effective items and tasks are deposited in the file. At first, it seems to be additional work, with very little return. However, in a few years, it is possible to start using some of the items and tasks from the file and

supplementing these with newly constructed items and tasks. As the file grows, it becomes possible to select the majority of the items and tasks from the file for any given test or assessment without repeating them too frequently. To prevent using a test item or assessment task too often, record the date it is used.

A file of effective items and tasks assumes increasing importance as we shift from test items that measure knowledge of facts to items and tasks that measure understanding, application, and thinking skills. Items and tasks in these areas are difficult and time consuming to construct. With all the other demands on our time, it is nearly impossible to construct effective test items or assessment tasks in these areas each time we prepare a new test or assessment. We have two alternatives: Either we neglect the measurement of learning outcomes in these areas (which, unfortunately, has been the practice), or we slowly build a file of effective items and tasks in these areas. If quality of student learning is our major concern, the choice is obvious.

SUMMARY

Some of the topics considered in this chapter are more relevant to traditional classroom tests involving right–wrong or single-best-answer items than to complex, performance-based tasks involving extended responses and more complicated scoring procedures. Test assembly and item analysis, for example, are more relevant for a classroom test containing many items than for a performance assessment involving a single task for a class period. At a conceptual level, however, the general principles considered in this chapter apply to complex performance-based assessments as well as to classroom tests involving only objective items.

The same care that goes into the construction of individual test items and assessment tasks should be carried over into the final stages of development and use. Attending to

Item Banking by Computer

Some schools use computers to maintain systematic item files (or item pools) for each of the various subjects and grade levels. The items are coded and stored by the test builder for easy retrieval. The code includes such things as instructional level, subject area, instructional objective, content topic, and item statistics (e.g., difficulty and discrimination indexes). This makes it possible to select items and build a test that matches a particular set of test specifications. The coded information concerning each item also aids in arranging the items in the test (e.g., by objective or order of difficulty).

The computer will print out these custom-designed tests and will also score, report, and analyze them. For examples and additional information, see the following Web sites:

Assessment System Corporation

http://www.assess.com

ERIC Clearinghouse on Assessment and Evaluation

http://ericae.net

Computer item banks are like any other item pool—you get out only what you put in. If you store ineffective items, you will get back ineffective items. Thus, item banking by computer requires careful screening of the items before they are entered.

the procedures for assembling, administering, scoring, and appraising the results will increase assurance that results are valid.

The preliminary steps in preparing the test will be simpler if items are recorded on cards. This facilitates the task of editing and arranging the items. Editing includes checking each item to make certain that its format is appropriate, that it is relevant to the specific learning outcome it measures, and that it is free from ambiguity, irrelevant clues, and nonfunctioning material. The final group of items selected for the test also should be checked against the table of specifications or other test plan to ensure that a representative sample of the learning outcomes and course content is being measured. In arranging the items in the test, all items of one type should be placed together in a separate section. The items within each section should be organized by the learning outcome measured and then placed in order of ascending difficulty.

The directions for the test or assessment should clearly convey the purpose of the measurement, the time allowed to finish, the basis for responding, and the procedure for recording the responses. The directions should indicate what to do about guessing for selection-type items. For performance-based tasks, the directions should describe the scoring procedure.

The procedures for administering the test or assessment should give all students a fair chance to demonstrate their achievement. Both the physical and the psychological atmosphere should be conducive to maximum performance. Unnecessary interruptions and unfair aid from other students or the teacher should be avoided.

Scoring the test can be facilitated by a scoring key or scoring stencil if separate answer sheets are used. Counting each right answer as 1 point is usually satisfactory. A correction for guessing is unnecessary on a typical classroom test for which students have sufficient time. Because assumptions underlying the use of correction-for-guessing formulas are debatable, it is recommended that they be used only with speeded tests. For most classroom tests, it is satisfactory to tell students to answer every question and then simply count the number of correct answers.

After the test has been scored, you should appraise the effectiveness of each item by means of item analysis. Use simple statistical procedures for determining the index of item difficulty (percentage of students who got the item right), item-discriminating power (the difference between high and low achievers), and the effectiveness of each distracter (degree to which it attracts more low achievers than high achievers). Item-analysis indexes can be computed quickly and easily if the data are based on the 10 highest-scoring and 10 lowest-scoring students (for class sizes ranging from 20 to 40 students). Because criterion-referenced mastery tests are designed to describe the learning tasks that students can perform rather than to discriminate among students, traditional indexes of item analysis are not used to select items for future tests—but they are relevant for detecting faulty items and for planning instruction. The results of item analysis are valuable in discussing the test with students, in planning remedial work, in improving teaching and testing skills, and in selecting and revising items for future use. Item-analysis data must always be interpreted cautiously because of their limited and tentative nature.

Building a file of effective test items and assessment tasks involves recording the items or tasks, adding information from analyses of student responses, and filing the records by both the content area and the objective that the item or task measures. Such a file is especially valuable in areas of complex achievement, when the construction of test items

and assessment tasks is difficult and time consuming. When enough high-quality items and tasks have been assembled, the burden of preparing tests and assessments is considerably lightened. Computer item banking makes the task even easier and is available in many schools.

LEARNING EXERCISES

1. What are the advantages of recording items during test construction?
2. List as many things as you can think of that might prevent a test item or assessment task from functioning as intended. Compare your list with the checklist on pages 352–353.
3. In what ways might poorly arranged items in a test adversely influence the validity of test results? What arrangement is best for valid results? Why?
4. What factors should be included in the general directions for a comprehensive departmental examination? How would the directions for a teacher's unit test differ?
5. What special precautions might be taken to avoid ambiguity, irrelevant clues, and other errors in objective test items?
6. Under what conditions should a correction for guessing be used to score a test?
7. If item-analysis data showed that an item was answered correctly by 7 of 10 students in the upper group and 3 of 10 students in the lower group, what would be the index of item difficulty? What would be the index of discriminating power? Would this item be considered effective or ineffective? Why?
8. How can you increase the discriminating power of a norm-referenced test?

REFERENCES

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Gronlund, N. E. (2005). *Assessment of student achievement* (8th ed.). Boston: Allyn & Bacon.

FURTHER READING

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Macmillan. Chapter 8, "Item Analysis," describes item-analysis procedures for norm-referenced tests with a brief introduction to item-response theory.

Baker, F. A. (1989). Computer technology in test construction and processing. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan. Describes the use of microcomputers for item writing, item banking, test construction, scoring, and reporting.

Crocker, L. (1992). Item analysis. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., vol. 1). New York: Macmillan. Provides an overview of item-analysis procedures and uses of the results.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice Hall. Chapter 13, "Using Item Analysis to Evaluate and Improve Test Quality," describes item analysis and illustrates its use in item revision.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Holt, Rinehart & Winston. Chapter 8, "Assembling, Reproducing, Administering, Scoring, and Analyzing Classroom Achievement Tests," presents a discussion of topics like those covered in this chapter.