PEARSON NEW INTERNATIONAL EDITION

Educational Assessment of Students

Anthony J. Nitko Susan M. Brookhart

Sixth Edition

ALWAYS LEARNING™

PEARSON®

Pearson New International Edition

Educational Assessment of Students

Anthony J. Nitko Susan M. Brookhart
Sixth Edition

PEARSON®

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsoned.co.uk

ISBN 10: 1-292-04102-1
ISBN 13: 978-1-292-04102-5

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Printed in the United States of America

Table of Contents

Glossary

absolute achievement: The achievement of specific learning targets and content without regard to what other students have achieved. See also **criterion-referencing** and **relative achievement**.

absolute standards grading: Evaluating student progress by comparing the student's achievement against content standards, performance standards, or learning targets rather than against the achievement of other students. See also **criterion-referencing**.

abstract/visual reasoning subtests: A subtest of a scholastic aptitude battery that contains items in which the examinee is expected to reason using geometric shapes, sequences, and patterns.

accountability testing: Assessment that is used to hold individual students or school officials responsible for ensuring that students meet state standards.

achievement: Knowledge, skills, and abilities that students have developed as a result of instruction.

activity-similarity rationale for assessing interests: Interest inventories built using this rationale present the students with lists of activities that are similar to those required of persons working in certain jobs or studying certain subjects.

adaptive assessment task: A computer-assisted assessment in which a student's response to one task will determine what the next presentation will be.

adaptive behavior: Behaviors indicating a child can cope with the normal social and physical environment that is appropriate for his or her age, especially outside the school context.

affective domain: A collection of educational outcomes and learning targets that focus on feelings, interests, attitudes, dispositions, and emotional states.

affective saliency: The degree of emotionality with which students hold particular attitudes.

age-based scores versus grade-based scores: *Age-based scores* use as a norm group only those students with the chronological age that is typical for a particular grade placement. *Grade-based scores* include all students at a particular grade placement, regardless of their chronological age.

algorithmic knowledge assessment: A diagnostic assessment that identifies whether a student knows the proper algorithm or procedure to follow to complete a given problem or task correctly. See also **component competencies of problem solving**.

alignment studies: Empirical studies involving the collection of ratings from trained judges and summaries of students' responses to testing. Their aim is to describe, as objectively as possible, the degree to which the actual test items on a state's assessment instrument(s) are matched to the educational content and performance standards set by that state.

"all of the above": A possible alternative or option for a multiple-choice item; all of the preceding alternatives are correct answers to this multiple-choice question.

alternate-forms reliability coefficient [delayed]: A procedure for estimating reliability that is used when one wants to study how scores are influenced by differences in both content and testing occasion. The procedure is to administer to the same group of students one form of an assessment on one occasion and an alternate form on another occasion.

alternate-forms reliability coefficient [same occasion]: A procedure for estimating reliability that is used when one wants to study the consistency of scores from two different, but comparable, samples of test items that were administered on the same occasion. The procedure is to administer two forms of an assessment to the same group of students on the same (or nearly the same)

From the Glossary of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

occasion and to correlate the scores. Also known as the *equivalent-forms reliability coefficient* or the *parallel-forms coefficient*.

alternate solution strategies: Different, but equally correct, procedures or methods for obtaining a correct solution to a problem or for producing the correct product.

alternative assessment: Generally refers to performance assessment. The "alternative" in alternative assessment usually means in opposition to standardized achievement tests and to multiple-choice (true-false, matching, completion) item formats. See also **performance assessment**.

alternatives: The list of choices from which an examinee answering a multiple-choice item must select the correct or best answer. Also known as *choices, options,* and *response choices*.

ambiguous alternatives: A type of multiple-choice and matching exercise item-writing flaw that results in *upper group students* being unable to distinguish between the correct answer and one or more of the distractors in a multiple-choice item.

analysis: A category in the Bloom et al. (1956) *Taxonomy of Educational Objectives*. Learning targets in this category ask students to identify the parts of a piece of information, explain the interconnections and relationships among the parts, or explain the organization or structure of the piece of information. See also **application, comprehension, evaluation, knowledge,** and **synthesis**.

analytic rubric, analytic scoring rubric: A rule that you use to rate or score the separate parts or traits (dimensions) of a student's product or process first, then sum these part scores to obtain a total score. See also **holistic rubric** and **rubrics**.

annotated holistic rubrics: Rules you use to conduct holistic rating of a student's product or process, then rate or describe a few characteristics that are strengths and weaknesses to support your holistic rating. See also **rubrics**.

annotated holistic scoring rubric: See **annotated holistic rubrics**.

application: A category in the Bloom et al. (1956) *Taxonomy of Educational Objectives*. Learning targets in this category ask students to use their knowledge and skills to solve new problems or to work effectively in new situations. See also **analysis, comprehension, evaluation, knowledge,** and **synthesis**.

aptitude: An aptitude for *X* is the present state of a person that indicates the person's expected future performance in *X* if the conditions of the past and present continue into the future (see Carroll, 1974).

area under the normal curve: The area in a segment of a normal curve between the graph of the curve and the horizontal axis.

argument-based approach to validation: Organizing the information used to demonstrate the validity of your assessment practices in the form of a persuasive argument. In other words, using a combination of logic and data to convince others that your interpretations and uses of the assessment results are valid. See also **assessment practice** and **validity**.

assessment: The process for obtaining information that is used for making decisions about students, curricula and programs, and educational policy. See also **evaluation, measurement,** and **test**.

assessment accommodations: See **assessment modifications**.

assessment modifications: Changes in either the conditions or materials of assessment that allow students with disabilities to be assessed in the same areas as students who are assessed with unmodified assessments.

assessment or test bias: See **bias (assessment or test)**.

assessment or test fairness: See **fair assessment or test**.

assessment planning: Planning ahead for what you will assess, when you will assess, how you will assess, how you will use the results of each assessment to guide your teaching, and the weight you will assign to each assessment.

assessment practice: When used in connection with assessment validation, it means the way an individual intends to interpret and use the assessment results in a particular situation. See also **argument-based approach to validation** and **validity**.

assessment variables: Characteristics about which you gather information needed for teaching, including sizing up the class and diagnosing students' needs, prerequisite achievements, attitudes, work habits, study skills, and their motivation and effort in school.

association test of personality: Examinees are presented with vague pictures as stimulus materials and asked to tell what their meaning is.

association variety of short answer: This format consists of a list of terms or a picture for which students have to recall numbers, labels, symbols, or other terms and write them next to the listed terms or given picture in the spaces indicated.

attitudes: A person's positive or negative feelings toward particular objects, situations, institutions, persons, or ideas. See also **interests** and **values**.

authentic assessment: A type of performance assessment in which students are presented with educational tasks that are directly meaningful instead of indirectly meaningful.

behavior checklist: A list of discrete behaviors related to a specific area of a student's performance that is used by observing a student and marking the behaviors on the list that were observed. See also **checklist**.

bell-shaped distribution: A distribution of scores that is unimodal, symmetrical, and the graph of which has the appearance of the cross section of a bell.

best answer item: A type of multiple choice item where all the distractors contain degrees of correctness, but one is best.

best works portfolio: A portfolio containing only a student's best final products or work in a subject. See also **portfolio**.

bias (assessment or test): A general term to describe a test or an assessment used unfairly against a particular group of persons for a particular purpose or decision.

bias as content/experience differential: According to this approach, an assessment usage is biased if the content of the assessment tasks is radically different from a particular subgroup of students' life experiences but the assessment results are interpreted without taking such differences into proper consideration.

bias as differential validity: According to this approach, an assessment would be biased if it predicted criterion scores better for one group of persons (e.g., whites) than for another (e.g., African Americans) (Cole & Moss, 1989).

bias as mean differences: According to this approach, an assessment usage is biased against a particular group when the average (mean) score of that group is lower than the average score of another group.

bias as misinterpretation of scores: According to this approach, an assessment usage is biased if someone who uses the results tries to make inappropriate inferences about students' performances that go beyond the content domain of the assessment (Cole & Moss, 1989).

bias as the statistical model: According to this approach, an assessment usage is biased if the statistical procedure used for selection is unfair to persons who are members of a particular group.

bias as the wrong criterion: According to this approach, an assessment usage is biased if the criterion measure that the test tries to predict is biased, making the selection process biased, even if the test is unbiased.

bias stemming from testing conditions: According to this approach, an assessment usage is biased if the scores are interpreted without considering that the basic stresses of test taking, such as test anxiety, feeling unwelcome, or being tested by a member of the opposite gender or another race, can adversely affect the performance of some groups.

bimodal distribution: A frequency distribution of scores in which there are two pileups of scores in two separate intervals. Sometimes referred to as a *U-shaped distribution*. See also **unimodal distribution**.

blueprint: See **table of specifications**.

borderline cases: Students whose marks place them at or very near the border between two letter grades.

carryover effect: A type of scoring error that occurs when your judgment of a student's response to Question 1 affects your judgment of the student's response to Question 2.

central tendency error: A type of error in rating students that occurs when a teacher fails to use extreme ratings and uses ratings in the middle part of the scale only.

checklist: A list of specific behaviors, characteristics of a product, or activities, and a place for marking whether each is present or absent.

checklist method of reporting student progress: A student progress report that contains a list of many specific behaviors, which a teacher checks off or rates as a student achieves each during the year.

choices: See **alternatives**.

clang association: A word in the stem of a multiple-choice or matching item that sounds very much like a word in the correct alternative.

classification decision: A decision that results in a person being assigned to one of several different but unordered categories, jobs, or programs. For example, children with disabilities may be classified into one (or more) of a few designated categories. See also **placement decision** and **selection decision**.

classification variety of multiple choice items: See **masterlist variety of matching exercise items**.

classroom activity versus assessments tasks: A performance activity used primarily as a teaching aid and without criteria for evaluating students' achievement is a *classroom activity* rather than an assessment task. *Assessments* must include achievement criteria and scoring guides in addition to a performance activity.

closed-response task: An assessment task allowing only a single correct answer. See also **open-response task**.

cloze reading exercise: A reading comprehension assessment method crafted by replacing every fifth word (excluding verbs, conjunctions, and articles) with blanks of equal length. The students' task is to read the passage and put in the missing words.

clueing: An item-writing flaw in which hints to the correct answer to one item are found in the contents of another item in the test.

***Code of Fair Testing Practices in Education (Revised)*:** A document based on the *Standards for Educational and Psychological Testing* that describes in nontechnical terms the obligations of test developers and test users to ensure that tests are used properly. See Appendix B.

***Code of Professional Responsibilities in Educational Measurement (CPR)*:** A document that describes the professional and ethical behaviors of test users regarding the interpretation, handling, and usage of test scores and materials. See Appendix C.

coefficient alpha reliability: A general version of the Kuder-Richardson formula 20 reliability coefficient. It is primarily used with tests and questionnaires that contain items that are scored more continuously.

cognitive domain: A collection of educational outcomes and learning targets that focus on a student's knowledge and abilities requiring memory, thinking, and reasoning processes.

combined group and individual project: A project that involves both a group activity and individual assessment components.

complete ordering of students: Ranking all students in the class on their test performance. See also **partial ordering of students**.

completion test of personality: A type of personality test that asks the examinee to complete sentences related to various aspects of self and of interpersonal relations (e.g., "Compared with most families, mine . . .").

completion variety of short answer: A format of achievement assessment that presents a student with incomplete sentences and asks the student to add one or more words to complete them correctly.

component competencies of problem solving: The linguistic, schematic, strategic, and algorithmic of subskills necessary to solve word problems in social studies, mathematics, and science. See also **algorithmic knowledge assessment, linguistic knowledge assessment, schematic knowledge assessment,** and **strategic knowledge assessment**.

comprehension: A category in the Bloom et al. (1956) *Taxonomy of Educational Objectives*. Learning targets in this category ask students to paraphrase or explain concepts in their own words. See also **application, analysis, evaluation, knowledge,** and **synthesis**.

concept: A name that represents a category of things such as persons, objects, events, or relationships.

concept mapping: A graphical way to represent how a student understands relationships among the major concepts in a subject.

concrete concept: A class, whose members have in common one or more physical, tangible qualities that can be heard, seen, tasted, felt, or smelled.

concurrent validity evidence: The extent to which individuals' current status on a criterion can be estimated from their current performance on an assessment instrument. See also **predictive validity evidence**.

confidence interval: See **uncertainty interval.**

confidentiality: The right of students to have their test results kept private, known only to authorized persons, and not released to those outside the school except by student-approved release.

content analysis of the responses: An analysis of students' responses to written open-ended items. Responses are organized into meaningful categories to identify in the class common errors and misconceptions.

content-based method for grading: See **quality-level method for grading**.

content centered: A criterion for a well-stated learning target: A learning target should describe the specific subject-matter content to which a student should apply the performance learned. See also **performance centered** and **student centered**.

content relevance: Validity evidence that focuses on whether the assessment tasks belong in the user's definition of what the assessment should include.

content representativeness: Validity evidence that focuses on whether the assessment tasks are representative of a larger domain of performance.

content standards: Statements about the subject-matter facts, concepts, principles, and so on that students are expected to learn.

context-dependent item sets: See **interpretive exercises**.

context-dependent tasks: See **interpretive exercises**.

continuous assessment: The daily process by which you gather information about students' progress in achieving the curriculum's learning targets (Nitko, 1995).

correct-answer variety: A multiple-choice item format in which one of the alternatives is unarguably the correct answer to the question or problem posed by the stem.

correction for guessing formulas: An algebraic formula used with response-choice items for adjusting each student's raw score by estimating how many items on which the examinee guessed. See also **response-choice items**.

correction variety of true-false items: An item format that requires students to judge a proposition, as does the true-false variety, but students are also required to correct any false statement to make it true.

correlation coefficient: A statistical index that quantifies, on a scale of −1 to +1, the degree of relationship between the scores from one assessment and the scores from another.

credentialing: Decision processes in which persons who meet specified requirements (usually involving passing a test or other assessment) are awarded a certain status and given a credential.

criteria for evaluating a planned assessment: These include (a) matching tasks to learning targets, (b) covering important skills, (c) selecting appropriate assessment task formats, (d) making assessments understandable, (e) satisfying validity criteria, (f) using the appropriate length of the assessment, (g) ensuring equivalence, and (h) identifying appropriate complexity and difficulty of tasks.

criterion-referenced grading framework: The assignment of grades by comparing a student's performance to a defined set of standards to be achieved, targets to be learned, or knowledge to be acquired.

criterion-referencing: A score-interpreting framework that compares a student's test performance against the domain of performances that the assessment samples to answer the question, "How much of the targeted learning did this student achieve?"

critical thinking: This is "reasonable, reflective thinking that is focused on deciding what to believe or do" (Ennis, 1985, p. 54). Critical-thinking educational goals focus on developing students who are fair-minded, are objective, reach sound conclusions, and are disposed toward seeking clarity and accuracy (Marzano et al., 1988).

curricular relevance: Validity evidence that focuses on the degree of overlap between a specific curriculum and the specific tasks on a particular assessment.

deadwood alternative: An alternative of a multiple-choice or matching item that no examinee chooses and, hence, is nonfunctional.

debate: A special type of oral performance that pits one student against another to logically argue issues. Assessment focuses on the logical and persuasive quality of the argument and the rebuttals.

decision consistency index: A statistical index used to describe the consistency of decisions made from scores rather than the consistency of the scores themselves.

decontextualized knowledge: The use of knowledge without its real-world context or application.

deficits in learning: Student learning needs identified by diagnostic assessment. Depending on the diagnostic approach used, deficits in learning can be low standing, relative to peers, on certain content or low performance on measures of prerequisite skills needed for particular learning targets.

defined concept: The name of a class, the members of which can be defined in the same way by attributes that are not tangible and which frequently involve relationships among other concepts. Sometimes called *abstract* or *relational concepts* (Gagné, 1970). See also **concept** and **concrete concept**.

demonstration: An on-demand complex performance assessment in which a student shows he or she can use knowledge and skills to complete a well-defined, "right or best way to do it" task.

derived score: A score obtained by statistically transforming a raw score in a way that increases its norm-referenced meaning. See also **percentile rank, raw score,** and **standard score**.

descriptive graphic rating scale: An improved version of the graphic rating scale in which the ambiguous single words (e.g., *frequently*) are replaced with short `behavioral descriptions of the various points along the scale. See also **rating scale** and **graphic rating scale**.

developmental learning targets: Skills and abilities that are continuously developed throughout life. Learning targets such as these are more aptly stated at a somewhat higher level of abstraction than mastery learning targets. See also **mastery learning targets**.

diagnostic assessment: Assessment of a student's learning difficulties that serves two related purposes: (a) to identify which learning targets a student has not mastered and (b) to suggest possible causes or reasons why the student has not mastered the learning targets.

dichotomous item scoring: Scoring an item in such a way that there are only two possible scores.

differential item functioning (DIF): An approach to studying item fairness at the level of individual test items rather than looking simply at average differences in an item's performance. The approach studies whether persons of the same ability, but from two different groups, performed differently on the item.

differentiated instruction: Using different instructional practices to meet the needs, abilities, interests, motivations of students, regardless of differences in ability. Characterized by clearly focused learning goals, pre-assessment and responses, flexible grouping, appropriate student choice during instruction, and ongoing formative assessment, differentiated instruction gives all students avenues to learning.

***DIQ*-score:** A type of normalized standard score, called a deviation IQ score. The distribution of such scores has a mean of 100 and a standard deviation of 15 (or 16 in some tests).

direct assessment: An assessment procedure that allows a student's learning target achievement to be expressed directly as it is intended by the learning target.

direction and intensity of attitude: Two of the ways students' attitudes may differ. Direction means that the attitude may be positive or negative. Intensity refers to the strength of feeling of a student's attitude. See also **attitudes**.

directions for matching: The instructions for a matching exercise that explain the basis the student is expected to use in matching the premises with the correct responses.

disaggregation of test results: Separation of test results for the total population of students in order to report on individual subgroups of students such as students who are poor or minorities, students with limited English proficiency, and students with disabilities, in addition to reporting on the total population.

dispositions toward critical thinking: The tendencies or habitual uses of critical thinking abilities.

distractor rationale taxonomy: A way to categorize the developmental level represented by an incorrect response choice in order to obtain diagnostic information from students' wrong answers to multiple-choice items (King, Gardner, Zucker, & Jorgensen, 2004).

distractors: Alternatives in a multiple-choice item that are not the correct or best answer to the question or problem posed by the stem, but that appear to be correct or plausible answers to less knowledgeable examinees.

domain of achievement: A description of all possible tasks that might be appropriate for assessing achievement for a particular set of learning targets.

dramatization: A performance assessment method that combines verbalizations, oral and elocution skills, and movement performances.

due process: Whether a person was treated fairly before a judgment was made against the person. *Substantive due process* concerns the appropriateness of the requirement (e.g., passing the test) and the purpose (e.g., maintaining high-quality teaching). *Procedural due process* focuses on the fairness with which the examinee was treated in the proceeding that led to a judgment.

educational goals: Statements of "those human activities which contribute to the functioning of a society (including the functioning of an individual *in* society), and which can be acquired through learning" (Gagné, Briggs, & Wager, 1988, p. 39).

electronic portfolio: A portfolio that is crafted as an electronic folder on a computer, a removable disk, or on a Web page. Entries are digitized files and may include text documents, photos, video clips, and other digitized material.

elements of a complete test plan: These include (a) content topics to assess, (b) types of thinking skills to assess, (c) specific learning targets to assess, and (d) emphasis (number of tasks or points) for each learning target to be assessed. See also **table of specifications**.

empirically documented tests: Refers to standardized tests that have data to support the development of the test, the selection of items, claims to reliability, and claims to validity.

empirically keyed scales: Interest inventory scales made up of items especially selected because research has shown that persons' responses to these items clearly differentiate between those who are currently and happily employed in a particular occupation and people in general.

empirical norming dates: Refers to the dates during the school year when tests were actually administered to students during the process of developing the grade-equivalent scores.

enhanced multiple-choice items: Multiple-choice items that assess combinations of skills and knowledge in ways that require students to apply what they know.

equivalence: In the context of classroom assessment, the degree to which past and present students are required to know and perform tasks of similar (but not identical) complexity and difficulty to get the same grade on the same content of the units. See also **table of specifications**.

error score: The score an examinee would obtain if you could quantify the amount of error in the examinee's obtained score. See also **obtained score** and **true score**.

ethnic and gender stereotyping: Depiction of races or genders in assessment material that is subtly or blatantly offensive to any subgroup of students or depicts races or genders in oversimplified inappropriate ways. See also **role stereotype.**

***ETS Test Collection*:** Educational Testing Service database of approximately 20,000 tests and other assessment instruments. It contains information on both published and unpublished instruments.

evaluation: The process of making value judgments about the worth of a student's product or performance. Also, a category in the Bloom et al. (1956) *Taxonomy of Educational Objectives*. See also **application, analysis, comprehension, knowledge,** and **synthesis**.

evaluation variables: See **assessment variables**.

exemplar: Examples of student work that illustrate or exemplify different levels on a scoring rubric.

expectancy table: A grid or two-way table that shows how criterion scores are related to test scores. It describes how likely it is for a person with a specific score to attain each criterion score level.

experiment: An on-demand performance used to assess a student's ability to plan, conduct, and interpret the results of an empirical research study that focuses on answering specific research questions or on investigating specific research hypotheses.

experiment-interpretation items: A type of context-dependent exercise in which an experiment and its results are presented to the examinee and the examinee must explain the results or choose the correct explanation from among several explanations presented.

expository writing: A type of writing that has as its purpose to give an explanation to and information for the reader.

expressed interests: What students will tell you their career or occupational interests are when you ask them directly. See also **interests, inventoried interests, manifested interests,** and **tested interests**.

extended normalized standard score: A type of normalized standard score that tells the location of a raw score on an achievement scale that spans multiple grades anchored to a lower grade reference group.

extended response essay item: A type of essay question that requires students to write essays in which they are free to express their own ideas, to show interrelationships among their ideas, and to organize their own answers. Usually no single answer is considered correct.

external assessment procedure: An assessment procedure that comes from outside the local school district and was not crafted by teachers in the district. A state's assessment and a standardized test are two examples.

external structure: Validity evidence that focuses on the pattern of relationships between assessment scores and external variables or criteria. See also **internal structure**.

extrapolation: The process of estimating an unknown number that lies outside the range of available data. Used extensively with standardized achievement tests to estimate the grade-equivalent scores of examinees whose raw scores lie well below or above the available data.

facial bias: According to this approach, an assessment is biased if it contains offensive stereotypes in its use of language or in pictures in the assessment tasks and materials (Cole & Nitko, 1981).

fair assessment or test: An assessment or test that provides scores that (a) are interpreted and used appropriately for specific purposes, (b) do not have negative or adverse consequences as a result of the way they are interpreted or used, and (c) promote appropriate values.

feedback to students: Information about how a student can improve his or her work, usually given by a teacher to a student on the basis of observation and diagnosis of performance on *formative assessments* or *classroom activities*. See also **formative uses of assessments**.

figural reasoning: The ability to reason using geometric figures: to infer relationships among the figures, to identify the similarities and differences among figures, and to identify progressions and predict the next figure in the progression.

filler alternative: A type of multiple-choice item-writing flaw in which a nonplausible alternative is added to an item primarily for the purpose of increasing the number of alternatives rather than as a useful functioning distractor.

fixed-percentage method for grading: Assigning grades by using percentages as bases for marking and grading papers. The relationship between percentage correct and letter grade is arbitrary.

foils: See **distractors**.

forced-choice item format: A technique used by interest inventories that presents activities in items in sets of three (triads). These items ask the student to mark the one activity in the triad that the student most ("M") likes and the one activity the student least ("L") likes. This is equivalent to asking a student to rank the three activities from most liked to least liked.

formative evaluation of schools, programs, or materials: Judgment about the worth of curricula, materials, and programs made while they are under development leading to suggestions for ways to redesign, refine, or improve them. See also **summative evaluation of schools, programs, or materials**.

formative evaluation of students' achievement: Judgment about the quality of students' achievement made while the students are still in the process of learning. Such judgments help you guide a student's next learning steps. See also **summative evaluation of students' achievement**.

formative uses of assessments: Using assessment results to improve your teaching and to help you guide students' learning. See also **summative evaluation of students' achievement**.

four principles for validation: Rules about interpretations, uses, values, and consequences that help you judge whether your assessment results have sufficient validity for their intended purposes.

frequency distribution: A table that shows the number of persons in a group having each possible score.

frequency polygon: A line graph of a frequency distribution.

functional alternatives: Response choices in a multiple-choice item that work effectively as distractors or correct answers. If no examinee in the lower scoring group chooses a particular alternative, it is considered nonfunctional. See also **distractors**.

gender representation: The number and ways that males and females are discussed or pictured in test items. See **role stereotype**.

gender stereotype: See **role stereotype**.

generalizability of assessment results: Validity evidence that focuses on the extent to which students' scores on a test can be generalized to their performance on the broader curriculum of the school district or state.

general learning targets: Statements of expected learning outcomes derived from educational goals that are more specific than the goals but not specific enough to be useful as classroom learning targets. See also **educational goals** and **specific learning targets**.

general scoring rubric: Guideline for scoring that applies across many different tasks, not just to one specific task. It may be used in its generic format or serve as a general framework for developing more specific rubrics. Also called *generic scoring rubric*.

general versus specific intellectual skills: *General intellectual skills* are a student's overall abilities to engage successfully in academic learning in general or on the average. *Specific intellectual skills* are the student's abilities to engage successfully in learning one subject or one academic area.

gradebook program: A computer program combining a spreadsheet and database that allows you to enter students' names and grades and then automatically calculates averages and letter grades.

grade-equivalent score *(GE)*: A norm-referenced growth scale score that tells the grade placement at which a raw score is average. A grade-equivalent score is reported as a decimal fraction, such as 3.4. The whole number part of the score refers to a grade level, and the decimal part refers to a month of the school year within that grade level.

grade mean equivalent: A norm-referenced score that tells the grade placement of a group's average expanded scale score.

grading: The process of summing up students' achievement in a subject through the use of letters such as A, B, C, D, and F.

grading for summative purposes: Assigning grades for the purposes of providing you, other teachers, school officials, students, parents, postsecondary educational institutions, and potential employers with a report about how well a student has achieved the curriculum learning targets.

grading on a curve: A method for assigning grades that ranks students' marks from highest to lowest, and assigns grades (A, B, C, etc.) on the basis of this ranking.

grading variables: The subset of variables, selected from among all the reporting variables, on which you may base your grades (Frisbie & Waltman, 1992). You use the grading variables to describe a student's accomplishments in the subject. See also **reporting variables**.

grammatical clue: A type of multiple-choice and matching item flaw in which the correct grammatical relationship between words in the stem and words in the correct alternate clue the examinee as to which alternative is correct. Similarly, incorrect grammatical relationships between words in the stem and distractors clue the examinee that those distractors can be eliminated from consideration.

graphic rating scale: A rating scale that contains an unbroken line to represent the particular achievement dimension and on which you rate a student's performance or product. See also **rating scale** and **descriptive graphic rating scale**.

greater-less-same items: A multiple-choice-item format that presents an examinee with a pair of concepts, phrases, quantities, and so on that have a greater-than, same-as, or less-than relationship and requires the examinee to identify what that relationship is.

grouped frequency distribution: A table showing the number of persons having specified intervals of scores. Unlike a frequency distribution, a grouped frequency distribution organizes the score scale into intervals, and then displays the number of persons with scores in each interval. See also **frequency distribution**.

group project: A long-term performance activity that requires two or more students to work together. The major purpose of a group project *as an assessment technique* is to evaluate whether students can work together in cooperative and appropriate ways to create a high-quality product. See also **individual student project**.

growth portfolio: A portfolio containing a selection of a sequence of a student's work that demonstrates progress or development toward achieving the learning target(s). See also **portfolio** and **best works portfolio**.

halo effect: A type of error that occurs when a teacher's general impression of the student affects how the teacher rates the student on specific dimensions.

heterogeneous alternatives: A type of item-writing flaw in which one or more alternatives of a multiple-choice item or matching exercise do not belong to the same set of things. See also **homogeneous alternatives**.

heuristic: Any one of several general strategies that may help solve a given problem.

high-quality assessment information: Assessment information that has high validity for the decisions for which you want to use it.

high-stakes assessments (tests): Assessments (or tests) of which the results are used for decisions that result in serious consequences for school administrators, teachers, or students.

histogram: A bar graph of a frequency distribution in which each frequency is represented by a rectangle. See also **frequency distribution**.

holistic rubric, holistic scoring rubric: Rubric that requires a teacher to rate or score a student's product or process as a whole without first scoring parts or components separately. See also **analytic rubric** and **rubrics**.

homogeneous alternatives; homogeneous premises and responses: A desirable item-writing practice in which each alternative of a multiple-choice or matching exercise is a member of the same set of "things," *and* each alternative is appropriate to the question asked or problem posed by the stem or premises.

homogeneous tasks: All of those tasks in one assessment that measure the same trait or ability.

homogeneous versus heterogeneous test: All items on a *homogeneous* test will measure one ability, whereas the items on a *heterogeneous* test will assess a combination of abilities.

IDEAL problem solver: A way of organizing general problem-solving skills into a five-stage process (Bransford & Stein, 1984):

I	Identify the problem
D	Define and represent the problem
E	Explore possible strategies
A	Act on the strategies
L	Look back and evaluate the effects of your activities

identifying errors in performance: A diagnostic assessment approach that identifies a student's errors, rather than reporting only a number-right total score reflecting overall performance on a particular learning target.

ill-structured problem: A type of problem in which the problem-solver must (a) organize the information to understand it; (b) clarify the problem itself; (c) obtain all the information needed, which may not be immediately available; and (d) recognize that there may be several equally correct answers.

imaginative writing: A type of writing in which the writer describes something that did not, often could not, happen.

incomplete stem: A type of multiple-choice item-writing flaw in which the stem does not contain enough information for the examinee to know what question or problem the item poses.

independent scoring of essays: When two or more raters score the same student's essay responses without consulting or collaborating with each other.

indirect assessment: A type of assessment that assesses part of the entire learning target or assesses the learning target in a context that is not intended by the learning target. See also **direct assessment**.

individualized education program (IEP): An educational plan designed by a child study team (including a teacher) and agreed to by the student's parents or guardians describing what learning targets the student should attain, the time frame for attaining them, the proposed methods for attaining them, and the methods of evaluating the student's progress in achieving the learning targets.

individual student project: A long-term performance activity during which students work independently and that results in a product that is one student's work: a model, a functional object, a substantial report, or a collection. See also **group project.**

informal assessment techniques: Impromptu methods you use to gather information that guides and fine-tunes your thinking while you are teaching, to plan your next teaching activities, and to diagnose the causes of students' learning difficulties.

informed consent: Giving approval to release information or participate in an activity after understanding (1) the extent to which personal information will remain anonymous, (2) the extent participation is voluntary, (3) who (or what agency) is requesting the information and for what purpose, and (4) what will happen to the information after it is collected.

in-level versus out-of-level testing: *Out-of-level testing* is using a standardized test designed for a certain grade level with students above or below that level; *in-level testing* is using a standardized test designed for students at that grade level.

intensity of attitude: See **direction and intensity of attitude**.

interacting with others: A critical thinking strategy requiring the use of rhetorical devices to persuade, explain, or argue.

interests: A person's preferences for specific types of activities when he or she is not under external pressure. See also **attitudes** and **values**.

interim or benchmark assessments: Assessments administered periodically during the school year to evaluate students' knowledge and skills relative to academic standards in order to inform policy makers or educators at the classroom, school, or district level.

internal structure: Validity evidence that focuses on the interrelationships among the individual tasks (items) on an assessment, and the relationship between the individual tasks and the total scores. See also **external structure**.

interpolation: The process of finding an unknown number that is between two known numbers. Used extensively in estimating the grade-equivalent scores of students who are not tested on the empirical norming dates of a standardized achievement test.

interpretive exercises: A set of items or assessment tasks that require the student to use reading material, graphs, tables, pictures, or other material to answer the items. See also **interpretive materials**.

interpretive materials: The reading material, graphs, pictures, tables, or other material that accompany a set of items and that the examinee must use to answer the questions or problems posed by the item.

interquartile range: The difference between the third and fourth quartiles. It is the range spanned by the middle 50% of the scores. See also **quartiles**.

inter-rater reliability: A procedure for estimating reliability used when you want to study the extent to which a student would obtain the same score if a different teacher had scored the paper or rated the performance.

inventoried interests: Career and vocational interests that are identified through various paper-and-pencil tests or interest inventories. See also **interests, expressed interests, manifested interests,** and **tested interests**.

IRT pattern score: A norm-referenced expanded-scale score derived from a mathematical equation that is fit to the publisher's sample of students' item responses. *IRT* stands for *item response theory*.

item analysis: The process of collecting, summarizing, and using information from students' item responses to make decisions about how each item is functioning.

item bank: A file of previously used items, usually along with the statistics about each item, that can be drawn upon to create new tests.

item difficulty index (*p* and *p):** The fraction of the total group answering a dichotomously scored item correctly. The item difficulty for a constructed-response and performance item, denoted *p**, is simply the average score for the group for that item.

item difficulty level: See **item difficulty index (*p* and *p**)**.

item discrimination index (*D* and *D):** For dichotomously scored items, *D* is the difference between the fraction of the upper group answering the item correctly and the fraction of the lower group answering it correctly. The discrimination index describes the extent to which a particular test item is able to differentiate the higher scoring students from the lower scoring students. See also **upper-, middle-, and lower-scoring groups**.

item response theory (IRT) score: See **IRT pattern score**.

key: The correct answer to any type of item or assessment task.

keyed alternative: The alternative in a multiple-choice or true-false item that is correct.

keyed answer: See **keyed alternative**.

keylist variety of matching exercise items: See **masterlist variety of matching exercise items**.

knowledge: A category in the Bloom et al. (1956) *Taxonomy of Educational Objectives*. Learning targets in this category ask students to recall information about facts, generalizations, processes and methods of doing things, theories, and so on. See also **application, analysis, comprehension, evaluation,** and **synthesis**.

knowledge structure assessment: A diagnostic assessment approach that identifies how a student (a) perceives the structure or organization of several concepts and facts of the subject, and (b) processes concepts and facts to solve problems in the subject.

Kuder-Richardson formula 20 reliability (*KR*20): A procedure for studying reliability when the focus is on consistency of scores on the same occasion and similar content, but when repeated testing or alternate forms testing are not possible. See also **coefficient alpha reliability**.

Kuder-Richardson formula 21 reliability (*KR*21): A procedure for studying reliability for the same purposes as Kuder-Richardson formula 20, except that this formula is used when the dichotomously scored test items are equally difficult, thus allowing for a simplified calculation procedure. See also **Kuder-Richardson formula 20 reliability (*KR*20)**.

learning hierarchy assessment: A diagnostic assessment approach that uses the prerequisite treelike ordering of learning targets to identify which learning targets a student has mastered and which have yet to be mastered.

learning objective: See **specific learning targets**.

learning progression: A description of development in a domain of knowledge along a continuum, usually with descriptions of what students at each level know or are able to do, often including misconceptions held at levels when knowledge is incomplete or not fully developed.

learning target: See **specific learning targets**.

leniency error: A type of rating error that occurs when a teacher tends to rate almost all students toward the high end of the scale and avoids using the low end. It is the opposite of a severity error. See also **severity error**.

letter grades method of reporting student progress: A summative evaluation of student achievement that uses letters (e.g., A, B, C, D, F) to describe achievement in each subject area.

letter to parents method of reporting student progress: A summative evaluation letter written by a teacher to describe a student's achievement in each subject area.

linear standard scores (*z, SS*): Norm-referenced scores that tell the location of the raw scores in relation to the mean and standard deviation of the distribution of all scores.

linguistic knowledge assessment: A diagnostic assessment approach that identifies the key terms and key phrases a student must understand to translate the problem statement into an internal model that can be solved. See also **component competencies of problem solving**.

linked items: An item-writing flaw in which the answer to one or more items depends on obtaining the correct answer to a previous item.

linking: See **linked items**.

local norm group: See **norm group (local, national, special)**.

local percentile rank: The percentile rank of a student in the distribution of scores for the school district the student attends.

logical error: A type of rating error that occurs when a teacher gives similar ratings on two or more dimensions of performance that the teacher believes are logically related but that are in fact unrelated.

logic rule method for grading: The use of a set of decision rules, based on student performance during a marking period, to assign grades. See also **quality-level method for grading** and **rubrics**.

mandated tests: Tests that students must take because the law says they are required to do so. State assessment programs are usually mandated.

manifested interests: Students' vocational and career interests that are deduced from what a student actually does, or the activities in which the student actually participates. See also **interests, expressed interests, inventoried interests,** and **tested interests**.

map-reading abilities: The abilities needed to obtain and use information from maps.

marking period: The period over which a teacher's summative evaluation of each student's achievement in each subject area is reported to the student, parents, and school officials.

masterlist variety of matching exercise items: A matching exercise that has three parts: (a) directions to students, (b) the masterlist of options, and (c) a list or set of stems.

mastery learning targets: Statements of what students can do at the end of instruction. Sometimes these are called "can do" statements (Forsyth, 1976), specific learning outcomes, or behavioral objectives. See also **developmental learning targets**.

mastery of specific objectives: A diagnostic assessment approach that identifies the specific learning targets a student has and has not mastered. See also **learning hierarchy assessment**.

matching exercise (basic): This format presents a student with three things: (a) directions for matching, (b) a list of premises, and (c) a list of responses. The student's task is to match each premise with one of the responses, using the criteria described in the directions as a basis for matching.

maximum performance assessment: Assessment of students when you set the conditions so that students are able to earn the best score they can. See also **typical performance assessment**.

MAZE item type: Reading comprehension assessment that is a multiple-choice adaptation of the cloze reading exercise. See also **cloze reading exercise**.

mean: An average score found by summing all of the scores and dividing by their number. Also known as the *arithmetic mean*.

measurement: A procedure for assigning numbers (usually called *scores*) to a specified attribute or characteristic of a person in such a way that the numbers describe the degree to which the person possesses the attribute. See also **assessment, evaluation,** and **test**.

measurement error: See **error score**.

median: The point on the score scale at which 50% of the scores are below and 50% are above.

median score method: A procedure for combining several component grades into a composite report card grade. All scores are converted to the same scale, usually a rubric or grade (A, B, C, D, F) scale, and the median mark is used as the composite grade.

mental age: The age at which the student's score on a scholastic aptitude test is average. This concept is no longer used in modern scholastic aptitude testing.

Mental Measurements Yearbook**s (MMYs):** A set of volumes published by the Buros Institute of Mental Measurement that contains reviews of tests published in the English language.

mental model: The way a person mentally represents or characterizes a problem before attempting to solve it.

metacognition: Knowledge of one's cognitive processes, including monitoring and regulating one's own learning.

minimum attainment method: A procedure for combining several component grades into a composite report card grade by the following process: determine which components of students' final grades are more important to demonstrating the students' achievement of the learning targets; specify, for each of these "more important" components, the minimum level of performance you will accept for each of the final grade levels; and establish rules for what levels of performance you will accept, at each final grade level, on each of the "less important" components. These rules form a set of decision rules for how to assign grades.

miskeyed items: Items for which the answer designated as correct in the answer key is wrong. An item may be miskeyed if a larger number of upper group students selects a particular wrong response.

modal-age norms: Norms that include, from among all students at a particular grade level, only those near the most typical chronological age for that grade.

mode: The most frequently occurring score in a distribution.

multilevel survey battery: A survey battery of standardized tests that spans a wide range of grades in each school subject. See also **single-level test**.

multiple-aptitude tests: Tests that assess several different abilities separately and provide an ability score for each. See also **omnibus test** and **two-score test**.

multiple-assessment strategy: Combining the results from several different types of assessments (such as homework, class performance, quizzes, projects, and tests) to improve the validity of your decisions about a student's attainments.

multiple-choice item: This item format consists of a stem that poses a question or sets a problem and a set of two or more response choices for answering the question or solving the problem. Only one of the response choices is the correct or best answer.

multiple marking system: A system of reporting summative evaluation of educational progress to students and parents using several kinds of symbols and marks. Multiple marking systems usually take the form of a report card and report on academic achievement, attendance, deportment, and nonacademic achievement.

multiple true-false variety of true-false items: This format looks similar to a multiple-choice item. However, instead of selecting one option as correct, the student treats every option as a separate true-false statement.

narrative report method of reporting student progress: A detailed, written report describing what each student has learned in relation to the school's curriculum framework and the student's effort in class.

narrative writing: A type of writing in which the author describes something that really happened, usually a personal experience of the writer.

national norm groups: See **norm group (local, national, special)**.

national percentile rank: A student's percentile rank in the national sample of students who took the test.

national stanines: Stanines are scores derived from a test publisher's national norm sample. See also **stanine scores**.

naturally occurring performance: Performance not done at the request of a teacher or school authority but which occurs in the normal course of daily activities.

NCE*-score:** See **normal curve equivalent *(NCE).

negative correlation: A type of relationship between two sets of scores that occurs when high scores on one assessment are associated with low scores on the other; low scores on one are associated with high scores on the other. See also **correlation coefficient** and **positive correlation**.

negatively discriminating item: An item that high-scoring students tend to answer incorrectly and low-scoring students tend to answer correctly.

negatively skewed distribution: A frequency distribution of scores in which the scores are piled up at the upper end of the score scale and spread thinly toward the lower end of the score scale. See also **positively skewed distribution**.

nondiscriminating item: An item for which the number of correct discriminations equals the number of incorrect discriminations (so that an equal number of upper and lower group students answers the item correctly).

"none of the above": A multiple-choice alternative that means that none of the preceding alternatives is the correct answer to the question or problem posed by the stem.

non-paper-and-pencil task: An assessment task in which performance is not primarily evaluated by the student's written response.

nonverbal tests: Tests that elicit and assess nonverbal responses such as assembling objects, completing experiments, performing a psychomotor activity, and so on. See also **performance assessment**.

normal curve equivalent *(NCE)*: A normalized standard score with a mean of 50 and a standard deviation of 21.06. This choice of standard deviation was made so the *NCE*-scores would span the range 1 to 99. It was developed primarily for use with federal program evaluation efforts (Tallmadge & Wood, 1976).

normal distributions: A set of theoretical distributions that takes on a bell-shaped and unimodal form through the use of a special mathematical formula.

normal growth (grade-equivalent view, percentile rank view): The *grade-equivalent view* of normal growth is that a student ought to exhibit a growth of 1.0 grade-equivalent unit from one grade to the next. Under this view, a student taking the test in second grade and scoring 1.3, for example, would need to score 2.3 in third grade, 4.3 in fifth, and so on to show "normal" or expected growth. The *percentile rank view* of normal growth is that a student shows normal growth if that student maintains the same position (i.e., percentile rank) in the norm from year to year.

normalized standard scores (*z_n, T, DIQ, NCE, SAT*): A category of scores in which the raw scores have been changed or transformed into other scores that are distributed more like a normal distribution.

normalizing a set of scores: The process used to transform the original raw scores in a distribution into a new set of scores that are distributed more like a normal distribution.

norm group (local, national, special): A well-defined group of students who have been given the same assessment under the same conditions (same time limits, directions, equipment and materials, etc.). See also **special norms**.

norm-referenced grading framework: A framework for assigning grades on the basis of how a student's performance (achievement) compares with other students in the class: Students performing better than most classmates receive the higher grades.

norm-referencing: A framework for interpreting a student's score by comparing his or her test performance with the performance of other students in a well-defined group who took the same test.

novel material: A new situation, problem, or context for applying previously learned knowledge or skills.

numbers method of reporting student progress: A summative evaluation of a student's achievement in each subject that is reported using either numbers (e.g., 5, 4, 3, 2, 1) or percentages.

numerical rating scale: A scale for which you must mentally translate judgments of quality or degree of achievement into numerical ratings.

objectivity: The degree to which two or more qualified evaluators of a student's performance will agree on what quality rating or score to assign to it.

obtained score: The scores students actually receive when you assess them. These scores include ratings from open-ended tasks such as essays, number-right scores from multiple-choice or short-answer tests, and standard scores or grade-equivalent scores from norm-referenced standardized tests. See also **error score** and **true score**.

odd-even split halves reliability coefficient: A procedure for estimating reliability when the focus is on consistency of scores on different samples of content on the same occasion, but when alternate forms have not been built. The items from one test are divided into two groups—the odd-numbered items in one group and the even-numbered in another. The full-test reliability is estimated from these two groups. See also

Spearman-Brown double length reliability formula and **split-halves reliability coefficient**

omnibus test: A reliability type of test containing items assessing several different abilities that comprise general scholastic aptitude, but that reports only a single score. See also **multiple-aptitude tests** and **two-score test**.

on-demand task: An assessment in which the teacher or other authority decides what and when materials should be used, specifies the instructions for performance, describes the kinds of outcomes toward which students should work, tells the students they are being assessed, and gives students opportunities to prepare themselves for the assessment.

open-response task: An assessment task allowing multiple correct answers. See also **closed-response task**.

optional essay questions: Presenting students with several different essays and allowing them to select which one(s) to answer.

options: See **alternatives**.

oral presentation: A performance assessment task that permits students to verbalize their knowledge and use their oral skills in the form of interviews, speeches, or other spoken activities.

overinterpreting score differences: Placing too much emphasis on small differences of students' obtained scores on a test or small differences in the obtained scores of one student on two different tests. See also **underinterpreting score differences**.

overlapping alternatives: A type of multiple-choice item-writing flaw in which the meaning of one alternative overlaps with or includes the meaning of another alternative.

paper-and-pencil assessments: Assessment techniques for which students write their responses to the questions. Written homework, seatwork, and tests are typical paper-and-pencil assessment techniques.

paper-and-pencil task: Assessment that requires students not only to record their answers but also to write explanations, articulate their reasoning, and express their own approaches toward solving a problem. Sometimes referred to as a *paper-and-pen* task.

parallel forms: Two forms (versions) of an assessment that are made up of tasks carefully matched to the same blueprint so the tests are as nearly alike as possible, even though they do not have any items in common.

parallel forms reliability coefficient: See **alternate forms reliability coefficient [same occasion]**.

parent-teacher conferences method of reporting student progress: A personal meeting between the parent(s) and the teacher that involves a summative report of a student's achievement in each subject.

partial credit: Giving the student some portion of an item's maximum possible points because the student's response is partially correct.

partial knowledge: The incomplete knowledge a student possesses and uses to respond to an item.

partial ordering of students: Placing students into two or more categories; the categories themselves are ordered, but there is *no ordering of individuals within a category*.

participation in assessment: Students with disabilities have the right, and sometimes the obligation, to be assessed, including taking part in accountability assessment programs.

passage dependency: The degree to which correct answers to questions on a reading comprehension test depend on the students actually reading and comprehending the passage.

passing score: The score that identifies students who have attained the minimum level of knowledge needed to benefit from further instruction on the topic. This may vary from one learning target to the next.

Pearson product-moment correlation coefficient: A type of correlation coefficient that is the average product of the linear z-scores corresponding to the paired scores in the set being correlated. It is denoted by ρ or r. See also **correlation coefficient**.

people-similarity rationale for assessing interests: The traditional view of describing a person's inventoried interests based on the rationale that "if a person likes the same things that people in a particular job like, the person will be satisfied with the job" (Cole & Hanson, 1975, p. 6).

percentage of agreement: An index of the consistency of decisions made by two independent judges. It is the percentage of students for whom the two judges reached the same decision.

percentages method of reporting student progress: A summative evaluation of a student's achievement in each subject that uses the average percentage of schoolwork marked correct.

percentile rank: A norm-referenced score that tells the percentage of persons in a norm group scoring lower than a particular raw score. See also **local percentile rank** and **national percentile rank**.

perfect matching: When a matching exercise has an equal number of premise statements and response statements.

performance assessment: Any assessment technique that requires students physically to carry out a complex, extended *process* (e.g., present an argument orally, play a musical piece, or climb a knotted rope) or produce an important *product* (e.g., write a poem, report on an experiment, or create a painting). The complexity of the task distinguishes performance assessments from the short answers, decontextualized math problems, or brief (one class period) essay tasks found on typical paper-and-pencil assessments.

performance centered: A criterion for a well-stated learning target: A learning target should describe what a student is able to do (or to perform) after completing instruction. See also **content centered** and **student centered**.

performance standards: Statements about the things students can perform or do once the content standards are learned. See also **content standards** and **standards**.

performance task: One activity or item in a performance assessment. See also **performance assessment**.

permanent record: The official summative record by grade level of a student's achievement in each subject and his or her attendance in a particular school.

personal bias: A type of rating error that occurs when a teacher has a general tendency to use inappropriate or irrelevant stereotypes favoring boys over girls, whites over blacks, working families over welfare recipients, or particular families and individual students a teacher likes over others the teacher may dislike.

persuasive writing: A type of writing in which the writer attempts to convince the reader of the writer's point of view. The writer may want the reader to accept his or her idea or to take some actions that the writer supports.

pictorial reasoning: The ability to reason using pictures. For example, to infer relationships among the pictured objects, to identify the similarities and differences among pictures, and to identify progressions and predict the next picture in the progression.

placement decision: A decision in which persons are assigned to different levels of the same general type of instruction, education, or work; no one is rejected, but all remain within the institution to be assigned to some level (Cronbach, 1990). See also **classification decision** and **selection decision**.

plausible distractor: An incorrect alternative of a multiple-choice or matching exercise that seems correct to less knowledgeable students.

poorly functioning distractor: A distractor in a multiple-choice item that virtually no one in the lower scoring group chooses.

portfolio: A limited collection of a student's work used for assessment purposes either to present the student's best work(s) or demonstrate the student's educational growth over a given time span.

portfolio culture model: An instructional approach advocating that students' portfolios become the center of a teacher's instructional planning and teaching activities so the teacher and the students will interact intensively with the portfolio contents (Duschl & Gitomer, 1991; Niyogi, 1995).

positive correlation: A type of relationship between two sets of scores that occurs when high scores on one assessment are associated with high scores on the other; low scores on one are associated with low scores on theother. See also **correlation coefficient** and **negative correlation**.

positively discriminating item: An item for which the proportion of upper scoring students getting high scores is larger than the proportion of lower scoring students getting high scores on it.

positively skewed distribution: A frequency distribution of scores in which the scores are piled up at the lower end of the score scale and spread thinly toward the upper end of the score scale. See also **negatively skewed distribution**.

positive or negative consequences of decisions: What happens to students as a result of taking an assessment. Positive consequences mean some desirable things happen (e.g., getting extra help in reading); negative consequences mean some undesirable things happen (e.g., being labeled as stupid because one needs extra help in reading).

predictive validity evidence: A type of external structure validity evidence showing the extent to which individuals' future performance on a criterion can be predicted from their prior performance on an assessment instrument. See also **concurrent validity evidence** and **external structure**.

preinstruction unit assessment framework: A plan to help assess cognitive and affective learning targets of an upcoming unit.

premises, premise list: The leftmost list of statements or elements in a matching exercise.

prerequisite knowledge and skill deficits assessment: A diagnostic assessment approach that identifies what a student needs to know before he or she can profit from new instruction. The approach uses task analysis to identify entry requirements and might also identify a learning hierarchy of prerequisites. See also **learning hierarchy assessment**.

prewriting activities: Before writing, a writer clarifies the purpose for writing, begins to organize thoughts, brainstorms, and tries out new ideas. The writer discusses the ideas with others, decides what the format and approach to writing will take, and determines the primary audience. A plan for the piece develops.

principle: A rule that describes what to do or the relationships between two concepts.

principle-governed thinking: Thinking that is manifested when a person consistently uses appropriate rules to identify how two or more concepts are related.

privacy: Keeping a student's assessment results closed to those who are unauthorized to have access to them. See also **confidentiality**.

problem: The presence of obstacles to attaining a desired outcome so that immediate attainment of a goal is not possible without further mental processing.

procedure checklist: A checklist of the steps necessary to complete a process correctly. See also **checklist**.

product checklist: A checklist of the necessary and important characteristics of the product a student is required to produce that is used to evaluate the quality of the work.

product versus process: The tangible thing a student produces is called a *product*. The procedure a student follows to complete a task or to produce the product is called a *process*.

professional responsibility: Acting toward students in a way that is ethical and consistent with one's role as a professional person.

progress monitoring: A method associated with Response to Intervention, using curriculum-based assessments to track and evaluate progress of students identified as at risk.

projective hypothesis: The assumption that an examinee's interpretations of vague stimuli (such as inkblots) will reveal the examinee's innermost needs, feelings, and conflicts, even though the examinee is unaware of what he or she is revealing (Frank, 1939).

projective personality test techniques: Assessment techniques that present the examinee with ambiguous stimuli (such as inkblots) and ask the examinee to respond to them.

prompt (or writing prompt): A brief statement that suggests a topic or question for students to write about, provides general guidance, motivates students to write, and elicits students' best writing performance.

proposition: Any sentence that can be said to be true or false. See also **true-false variety**.

psychometric issues: Issues about assessment, especially bias in assessment, that concern the technical or statistical properties of the assessment in question.

psychomotor domain: A collection of educational outcomes and learning targets that focus on motor skills and perceptual processes.

pupil-teacher conferences method of reporting student progress: A method of reporting a student's summative achievement evaluation by means of a direct meeting between the teacher and student.

purging records: Destroying recorded information no longer needed for making decisions about a student so persons who are unauthorized to have access to that information cannot use it.

quality-level method for grading: A method for assigning letter grades in which the type of student performance required for each letter grade is specified beforehand. See also **logic rule method for grading** and **rubrics**.

quantitative reasoning: Reasoning with numerical quantities. For example, to infer relationships among the numbers, to identify the similarities and differences among numbers and patterns, and to identify progressions and predict the next number in the progression.

quartiles: Points on the score scale that divide the group of scores into quarters.

random guessing: Responding to an item using chance rather than using your knowledge.

range: The difference between the highest and lowest scores in a set. It is used as a simple index of the spread of the scores in the set.

rater drift: A type of rating error that occurs when the raters, whose ratings originally agreed, begin to redefine the rubrics for themselves. As a result, the raters no longer produce ratings that agree.

rating scale: A scoring rubric that helps a teacher assess the degree to which students have attained the achievement dimensions in the performance task. See also **checklist**.

rating scale method of reporting student progress: A summative evaluation of a student's achievement that uses a rating scale to describe the degree of mastery. See also **rating scale**.

raw score: The number of points (marks) you assign to a student's performance on an assessment. Points may be assigned based on each task, or points awarded on separate parts of the assessment.

readiness test: An assessment of a student's general developmental skills needed for first-grade work, especially reading, where grouping by readiness level is a common practice.

recency of norm data: How current the norm data are. As the curriculum, schooling, and social and economic factors change, so will the currency of the data.

relational concepts: See **defined concept**.

relative achievement: The level of a student's achievement expressed in terms of comparisons to peers rather than by describing the specific learning targets the student has achieved. See also **absolute achievement** and **criterion-referencing**.

relative standards grading: See **norm-referenced grading framework**.

relevance of norm data: The extent to which the norm group a publisher provides is the appropriate group to which you want to compare your students' performance on the test.

reliability: The amount of consistency of assessment results (scores). Reliability is a limiting factor for validity.

reliability coefficient: Any of several statistical indices that quantifies the amount of consistency in assessment scores. See also **reliability**.

reliability decay: A rating error that results in the scores from multiple raters becoming less consistent over time.

reliability of ratings: The consistency of students' ratings over time, different samples of content, and different raters. See also **reliability**.

report card: The document that reports the summative achievement grades to students and parents.

reporting variables: A subset, from among all the assessment variables, that a school district will expect a teacher to report to parents and for official purposes (Frisbie & Waltman, 1992).

representativeness of norm data: The extent to which the norm sample is based on a carefully planned sample that represents the target population. The test publisher should provide you with information about the subclassifications (gender, age, socioeconomic level, etc.) used to ensure representativeness.

response-choice items: Test items that provide students with alternatives from which to choose to answer the question or solve the problem posed.

response list: The list of plausible response alternatives in a matching exercise. This list is placed to the right of the premise list when crafting exercises. See also **premises**.

Response to Intervention (RTI): An initiative that many states are using to identify students in need of special assistance and to provide tiers of assistance in order to minimize the number of students identified for special education services. RTI defines students who do not progress in otherwise effective instruction as not responsive to that instruction.

restricted-response essay items: Essay prompts or instructions that restrict or limit both the substantive content and the form of the written response.

right-wrong variety of true-false items: This item format presents a computation, equation, or language sentence that the student judges as correct or incorrect (right or wrong).

role stereotype: Depiction in assessment materials of races or genders in oversimplified activities or work roles that convey the impression that such persons' capacities are limited in some way.

rubric method for grading: See **logic rule method for grading** and **quality-level method for grading**.

rubrics: A coherent set of rules you use to evaluate the quality of a student's performance: They guide your judgments and ensure that you apply the rules consistently from one student to the next. See also **checklist** and **rating scale**.

***SAT*-score:** A normalized standard score from a distribution that has a mean of 500 and a standard deviation of 100.

scaffolding: The degree of support, guidance, and direction you provide students when they set out to complete the task.

scatter diagram (scattergram): A graph on which paired scores are plotted to show their relationship.

schema (schemata): The way knowledge is represented in a person's mind through networks of connected concepts, information, rules, problem-solving strategies, and conditions for actions.

schema-driven problem solving: When a person recognizes a particular problem as part of or very similar to an existing schema and applies the solution strategy stored in that schema to solve the new problem (Gick, 1986). See also **schema (schemata)**.

schematic knowledge assessment: A diagnostic assessment approach that identifies whether a student has formed an internal representation or model of the problem. See also **component competencies of problem solving**.

school averages norms: A tabulation of the average (mean) score from each school building in a national sample of schools that provides information on the relative ordering of these averages (means).

score band: See **uncertainty interval**.

scorer reliability: See **inter-rater reliability** and **decision consistency index**.

scoring key: A rubric or list of rules that shows the correct answer and the kinds of partially correct answers that are to receive various amounts of credit.

scoring rubric: See **rubrics**.

selection decision: A decision in which an institution or organization decides that some persons are acceptable whereas others are not; those unacceptable are rejected and are no longer the concern of the institution or organization. See also **classification decision** and **placement decision**.

self-evaluation checklist: A checklist that students use to evaluate their own performance.

self-referenced grading framework: The assignment of grades by comparing a student's performance with his or her own past performance or your perceptions of his or her capability.

severity error: A rating error that occurs when a teacher tends to assign almost all ratings toward the low end of the scale. It is the opposite of a leniency error. See also **leniency error**.

short-answer variety: This item format requires a student to respond with a word, short phrase, number, or symbol.

short-term memory subtests: Assessments of a person's ability to remember patterns, objects, words, and numbers immediately after they are heard, seen, or read.

simulation: On-demand event that happens under controlled conditions and that attempts to mimic naturally occurring events.

single-level test: A standardized survey battery that is used only at one grade level or one narrow range of grade levels. See also **multilevel survey battery**.

Six + 1 Traits® of Writing: A framework and scoring rubrics for assessing general writing ability that focuses on evaluating a student on seven writing traits for each essay: ideas, organization, voice, word choice, sentence fluency, conventions, and presentation.

sizing up: Using assessment information to form a general impression of a student's strengths, weaknesses, learning characteristics, and personality at the beginning of a course or of the year.

skewed distribution: A description of a frequency distribution in which the scores are piled up on one end of the score scale and thinly spread out toward the other. See also **negatively**.

SOAP: An acronym for the following elements that should appear in the prompt to stimulate good writing on the part of the student (Albertson, 1998):

S *Subject*—inform the student who or what the piece is supposed to be about.

O *Occasion*—inform the student what is the occasion or situation that requires that the piece be written.

A *Audience*—inform the student who the intended audience is.

P *Purpose*—inform the student what the purpose is supposed to be: Is it to inform or narrate? To be imaginative? To be persuasive?

Spearman-Brown double length reliability formula: A procedure for estimating reliability when the focus of the study is on consistency of students' scores from one sample of items to another equivalent sample of items from the same content domain, but when only one form of the test exists. See also **odd-even split halves reliability procedure** and **split-halves reliability coefficient**.

special norms: Percentile rank or standard-score norms developed for specific subpopulations of students such as students with hearing impairments, Catholic school students, and so on.

specific determiner: A word or phrase (e.g., *always, never, often, usually,* and so on) in a true-false or multiple-choice item that "overqualifies" a given statement and gives the student an unintended clue to the correct answer (Sarnacki, 1979).

specific learning targets: A clear statement about what students are to achieve by the end of a unit of instruction. See also also **educational goals** and **general learning targets**.

specimen set: A packet of materials from a test publisher containing a sample of the test, sample computer reports, promotional materials, and (occasionally) a technical report of the test's quality.

speeded assessment: Any assessment that focuses on how quickly a student can perform.

spiral format: An arrangement of items in a test whereby similar types of items are not grouped together into subtests, but are arranged in a pattern so that one item of each type is presented; then the sequence is repeated, but with more difficult items.

split-halves reliability coefficient: Any method for estimating reliability on a single occasion by studying the relationship between students' scores on each half of the full-length test. See also **domain of achievement, Spearman-Brown double length formula,** and **odd-even split-halves reliability procedure**.

***SS*-score:** A type of linear standard score that tells the location of a raw score in a distribution having a mean of 50 and a standard deviation of 10. See also **linear standard scores** and **raw score**.

***SS*-score method for making composites:** A method for preparing students' composite marks for purposes of norm-referenced grading that preserves the influence (weights) you want the components of the composite to have.

stability coefficient: Any of several methods for estimating reliability that study the consistency of students' scores from one occasion to the next. See also **alternate forms reliability coefficient [delayed], alternate forms reliability coefficient [same occasion],** and **test-retest reliability coefficient**.

stages in crafting performance tasks: Three stages of developing a performance task are: (a) being very clear about the performance you want to assess, (b) crafting the task, and (c) crafting a way to score and record the results (Stiggins, 1994).

stakeholders: Persons or groups with an interest in the results of an assessment, usually because they will be affected by decisions made about them using the test results.

standard age score (*SAS*): Normalized standard score with a mean of 50 and standard deviation of 8 in the norm group having the same age as the student being tested.

standard deviation: An index of the spread of the scores in a distribution calculated by taking the square root of the mean squared deviation of the scores from the arithmetic mean of the scores.

standard deviation method of grading: A norm-referenced grading method that uses the standard deviation of the class' scores as a unit of measure on the grading scale: A teacher computes the standard deviation of the scores and uses this number to mark off segments on the number line that define the boundaries for grade assignment. See also **standard deviation**.

standard error of measurement (*SEM*): An estimate of the standard deviation or the spread of a hypothetical obtained-score distribution resulting from repeated testing of the same person with the same assessment. See also **obtained score** and **standard deviation**.

standardized patient format: Originally used to assess the clinical skills of medical candidates and practicing doctors, an actor is trained to display the symptoms of a particular disorder. Each medical candidate meets and interviews this standardized patient to diagnose the illness and to prescribe treatment.

standardized test: A test for which the procedures, administration, materials, and scoring rules are fixed so that as far as possible the assessment is the same at different times and places.

standards: Statements about what students are expected to learn. Some states call these statements *essential skills, learning expectations, learning outcomes, achievement expectations,* or other names. Often there are two sets of standards: content and performance. See also **content standards** and **performance standards**.

standard score: A category of transformed scores that changes the mean, standard deviation, and sometimes the shape of the distribution of the original scores so they are more easily interpreted. See also **linear standard scores (*z,SS*)** and **normalized standard scores (z_n, *T, DIQ, NCE, SAT*)**.

Standards for Educational and Psychological Testing: Guidelines and recommendations prepared by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education for the development and use of educational and psychological assessments.

standards-referencing: A score-interpreting framework that compares a student's test performance to clearly defined levels of achievement of proficiency. These levels are established using both criterion-referencing and norm-referencing techniques. See also **criterion-referencing** and **norm-referencing**.

stanine scores: A type of normalized standard score that tells the location of a raw score in one of nine specific segments of a normal distribution. Thus, stanine is derived from standard nine.

state-mandated assessments: Tests and other assessments that the law requires to be administered to all students at designated grade levels.

statement-and-comment items: Items used to assess students' ability to evaluate a given set of interpretations of quoted comments using learned criteria.

statistic (statistical index): A summary number that concisely captures a specific feature of a group of scores.

stem: The part of a multiple-choice item that asks a question or poses a problem to be solved.

strategic knowledge assessment: An assessment approach that pinpoints a student's ability to identify the proper sequence of steps or the proper processes needed to reach the answer. See also **component competencies of problem solving**.

strip key: A strip of paper on which the correct answers to completion items are written in a column in such a manner that when the strip is put on a student's test paper, the correct answers line up with the locations of the student's responses.

structure a task: To provide written or oral guidance to a student for how to complete a task, what resources to use, how long the response must be, and so on. The more guidance you give, the more structured the task is said to be. See also **scaffolding**.

structured (self-report) personality assessment techniques: These assessment procedures have a specific set of response-choice items; follow very specific rules for administering, scoring, and interpreting the tests; and require examinees to respond to the items in a way that describes their personal feelings (e.g., examinees may be asked whether the statement, "I usually express my personal opinions to others," is true of themselves).

structured task (exercise): See **structure a task** and **scaffolding**.

student centered: A criterion for a well-stated learning target: A learning target should describe what a student is to learn. See also **performance centered** and **content centered**.

student progress reporting methods: Any one of several ways in which schools and teachers report each student's achievement to parents and for the official school records. These include letter grades, number grades, percentage grades, checklists, rating scales, narrative reports, pupil-teacher conferences, parent-teacher conferences, and letters to parents.

student self-assessment: Involving students in judging the quality of their own work against learning targets and in deciding what actions they need to take to improve. See also **formative uses of assessments**.

subtest: A short test that is scored separately but is part of a longer battery of tests. The longer battery is comprised of two or more subtests.

summative evaluation of schools, programs, or materials: Judgments about the worth of programs, curricula, or materials after they are completed with the idea of suggesting whether they should be adopted or used. See also **formative evaluation of schools, programs, or materials**.

summative evaluation of students' achievement: Judgments about the quality or worth of students' achievement after the instructional process is completed. See also **formative evaluation of students' achievement**.

summative uses of assessment: See **summative evaluation of students' achievement**.

surface feature: A diagnostic assessment approach that uses the immediate external feature of the content of a test or test item to describe a student's achievement. This is contrasted with the deeper features of how a student perceives the structure or organization of that content, and processes information and knowledge to solve problems using that content knowledge. See also **knowledge structure assessment**.

symmetrical distribution: A frequency distribution of scores in which it is possible for the graph of the distribution to be folded along a vertical line so that the two halves of the figure coincide

synthesis: A category in the Bloom et al. (1956) *Taxonomy of Educational Objectives*. Learning targets in this category ask students to combine parts into a whole that was not there before.

table of specifications: This chart describes the major content categories and skills that a test assesses. It describes the percentage of tasks (items) for each content-skills combination included on the test.

tabular (matrix) items: A type of matching exercise in which the student matches elements from several lists of *responses* (e.g., presidents, political parties, famous firsts, and important events) with elements from a common list of *premises*.

tandem arrangement of alternatives: A type of multiple-choice item-writing flaw in which the alternatives are arranged in a paragraph-like continuous stream of text instead of the more desirable list arrangement of one alternative placed beneath the other.

task-directed thoughts: Thoughts and test-taking actions that focus on completing the assessment tasks and thereby reduce any tensions that are associated with them (Mandler & Sarason, 1952). See also **task-irrelevant thoughts**.

task format: The way a task or item appears on an assessment. Typical formats include multiple-choice, true-false, matching, short-answer, and essays, among others.

task-irrelevant thoughts: Thoughts and test-taking actions that are self-preoccupied, centering on what could happen if a student fails a test or on a student's own helplessness, and sometimes on a desire to escape from the test situation as quickly as possible (Mandler & Sarason, 1952). One of the four test anxiety factors. See also **task-directed thoughts**.

task-specific rubrics: Scoring rubrics in which the description of quality levels refers to the specific task and expected responses. See also **general scoring rubric** and **rubrics**.

taxonomies of instructional learning targets: Highly organized schemes for classifying learning targets (instructional objectives) into various levels of complexity. See also **cognitive domain, affective domain,** and **psychomotor domain**.

teaching actions after assessing: The things you do to use the assessment results you obtain to improve your teaching and your students' learning.

technical manual: A publication prepared by a test developer that explains the technical details of how the test was developed, how the norms were created, the procedures for selecting test items, the procedures for equating different forms of the test, and reliability and validity studies that have been completed for the test.

test: An instrument or systematic procedure for observing and describing one or more characteristics of a student using either a numerical scale or a classification scheme. See also **assessment, evaluation,** and **measurement.**

test anxiety: Increased emotional tension among students who want to do well on a test that results in bodily and autonomic arousal and thoughts about the negative consequences of failure and how a student's performance will compare to others.

***Test Critiques*:** A series of volumes that reviews the most frequently used tests in business, education, and psychology. Published by the Test Corporation of America.

tested interests: Students' vocational and career interests inferred from the results of an assessment of a student's information and knowledge of a particular subject matter. See also **interests, expressed interests, inventoried interests,** and **manifested interests**.

test level: The grade level or narrow range of grade levels for which a standardized test is targeted.

test-retest reliability coefficient: A procedure for estimating reliability when the focus of the study is the consistency of the students' scores from one occasion to the next on the same test items.

***Tests in Microfiche*:** A collection of more than 800 unpublished tests used in education, business, and psychology. Published by the Educational Testing Service.

***Tests in Print*:** A test bibliography that contains information on more than 2,900 commercially available instruments. Published by the Buros Institute of Mental Measurements.

test-takers' rights: The rights of those who take tests to information from and fair treatment by those who administer tests and use the results.

testwiseness: A student's ability to use the characteristics of both the assessment materials and the assessment situation to attain a higher score than the student's knowledge would otherwise warrant.

total points method for grading: A criterion-referenced method of assigning grades in which each component included in the final composite grade is given maximum point value (e.g., quizzes may count 10 points, exams may count a maximum of 50 points each, and projects may count a maximum of 40 points each); letter grades are assigned on the basis of the number of total points a student accumulated over the marking period.

true-false variety: An item format consisting of a statement or proposition that the student must judge as true or false. See also **proposition**.

true score: The hypothetical score you would obtain if you subtracted the examinee's error score from the examinee's obtained score. See also **error score** and **obtained score**.

***T*-score:** A type of normalized standard score that tells the location of a raw score in a normal distribution having a mean of 50 and a standard deviation of 10. The normalized *T*-score is the counterpart to the linear *SS*-score.

two-category method of reporting student progress: A method for reporting summative evaluations of student achievement that uses only two levels of achievement such as pass-fail.

two-score test: A type of test that assesses several kinds of specific abilities, but reports only two scores, usually verbal/quantitative or verbal/nonverbal. See also **omnibus test** and **multiple-aptitude tests**.

types of test-anxious students: There are three types of test-anxious students: those who do not have good study skills and fail to understand how the main ideas of the subject you are teaching are related and organized; those who do have a good grasp of the material and good study skills but have built up fears of failure associated with assessment and evaluation; and those who believe they have good study habits but who do not.

typical performance assessment: Gathering information about what a student would do under ordinary or everyday conditions. See also **maximum performance assessment**.

uncertainty interval: The score interval within which an examinee's true score is likely to be. The endpoints of this interval are calculated by (a) subtracting the standard error of measurement from an examinee's obtained score (lower endpoint) and (b) adding the standard error of measurement to the examinee's obtained score (upper endpoint). Also referred to as the score band. See also **obtained score, standard error of measurement (*SEM*),** and **true score**.

underinterpreting score differences: A type of score-interpretation error that occurs when differences in scores between two students or differences in scores of one student on two tests are ignored even though the differences are not due simply to error of measurement. Some action should be taken. See also **overinterpreting score differences**.

unimodal distribution: A frequency distribution of scores in which there is one pileup of scores (i.e., one mode). See also **bimodal distribution**.

unit of instruction: A teaching sequence covering from 1 to 7 weeks of lessons, depending on the students and topics you are teaching.

universal design: A concept that originated in the field of architecture. In assessment, it means designing assessments to be accessible to as many students as possible, to the greatest extent possible, without the need for accommodations or modifications.

upper-, middle-, and lower-scoring groups: The three groups into which you divide the class before conducting an item analysis. The groups are formed after ranking students on the basis of their total score on the test that includes the items you will be analyzing.

validity: The soundness of your interpretations and uses of students' assessment results.

validity coefficient: A predictive or concurrent correlation that is used as one piece of external structure evidence to support the validity of an assessment. See also **correlation coefficient, concurrent validity evidence, external structure,** and **predictive validity evidence**.

values: A person's long-lasting beliefs of the importance of certain life goals, a lifestyle, a way of acting, or a way of life. See also **attitudes** and **interests**.

verbal clues: See **grammatical clue** and **specific determiner**.

verbal comprehension tests: Tests that assess the students' ability to understand verbal material and to use language to express themselves.

verbal reasoning tests: Tests that assess students' ability to see relationships among words, read critically, and reason with words.

vocational interest inventories: Formal paper-and-pencil questionnaires that help students express their likes and dislikes about a wide range of work and other activities. A pattern of vocational and career interests is then determined from the students' responses.

well-structured problems: Problems are presented as assessment tasks that are clearly laid out: All the information students need is given, the situations are very much the same as students were taught in class, and there is usually one correct answer that students can attain by applying a procedure that was taught (Frederiksen, 1984).

window dressing: The use of words that tend to "dress up" an item stem to make it sound as though it is testing something of practical importance, when it does not (Ebel, 1965).

writing process: Most writing results from an orderly process that includes drafting, feedback, revisions, and polishing.

writing traits (writing dimensions): The several characteristics or qualities that can be used to evaluate writing quality. Each characteristic is expressed as a continuum of quality. See also **Six + 1 Traits® of Writing**.

yes-no variety of true-false items: An item format that asks a direct question, to which a student's answer is limited to yes or no.

yes-no with explanation variety of true-false items: An item format that asks a direct question and requires the student to respond yes or no and explain why his or her choice is correct.

***z*-score:** A type of linear standard score that tells the number of standard deviation units a raw score is above or below the mean of a given distribution. The mean and standard deviation of the distribution of *z*-scores are always zero and one, respectively. See also **linear standard scores** and **raw score.**

z_n-score: A type of normalized standard score: z_n-scores have percentile ranks corresponding to what would be expected in a normal distribution. See also **normalized standard scores**.

Classroom Decision Making and Using Assessment

KEY CONCEPTS

1. Assessment provides teachers with information to make decisions about teaching and provides students with information to make decisions about learning.
2. *Assessment, test, measurement,* and *evaluation* are different but related terms.
3. High-stakes assessments provide those in authority with the information they use to classify and sanction.
4. Different kinds of educational decisions require different types of assessment information.
5. Professional guidelines for assessment competencies and assessment literacy are available.

IMPORTANT TERMS

accountability testing

assessment

classification decisions

content standards

credentialing

diagnostic assessments

disaggregation of test results

evaluation

formative evaluation of schools, programs, or materials

formative evaluation of students' achievement

high-stakes assessments (tests)

measurement

performance standards

placement decisions

selection decisions

summative evaluation of schools, programs, or materials

summative evaluation of students' achievement

test

From Chapter 1 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

ASSESSMENT AND CLASSROOM DECISIONS

It is almost impossible for you to have attended school without having been exposed to a wide variety of educational and psychological assessment procedures. The fact that you are reading this book for a testing and measurement course places you not only among test takers, but also among successful test takers. Think for a few minutes: How many ways have you been assessed in your life? When did your assessment experiences begin? Consider this example:

Example

Meghan's educational assessment began in kindergarten with an interview and an observation. The state in which she lived had no mandatory kindergarten requirement. On registration day, Meghan and her mother came to school and were interviewed briefly. A teacher rated Meghan's cognitive and social-emotional skills. Her development was judged normal, and she attended kindergarten.

During the year, she experienced difficulty in paying attention to the teacher and participating in group activities, although she was neither aggressive nor hostile. She was given a "readiness test" at the end of kindergarten and performed as an average child. Her teacher recommended that she continue on to first grade, but her parents balked: They didn't think she was ready.

They took her to a child guidance clinic and requested further psychological assessment. The clinical psychologist administered an individual intelligence test and a "projective test" in which Meghan was asked to tell a story about what was happening in each of a set of pictures. The psychologist interviewed her, her parents, and her teacher. The psychologist described her as normal, both in cognitive ability and in social-emotional development.

Her parents withdrew her from the school she was attending and placed her in another school to repeat kindergarten. Later, they reported that whereas her first experience was difficult for her, her second kindergarten year was a great success. In their view, a teacher who was particularly sensitive to Meghan's needs helped accelerate her cognitive development. By the end of the year she had also become more confident in herself and regularly participated in group activities.

This brief anecdote shows assessments being used early in the person's life. Most of us recall more easily the assessments applied to us later in our lives, as older children, and as adults. You may not even associate the term *assessment* with Meghan's interviews. Yet, as we explain later, interviews are included in the broad definition of assessments.

Meghan's situation also illustrates that assessment results can contribute to a decision, but everyone concerned may not interpret the results in the same way. Although Meghan's parents may have been right to have her repeat kindergarten, there is no way of knowing what would have happened had she gone straight to first grade, because she didn't.

Decisions involve using different kinds of information. Sometimes test scores play a major role; at other times, less formal assessments play a more dominant role. In Meghan's case, both informal (teachers' observations, interviews) and formal (readiness test, intelligence test, projective test) assessments were administered.

Making good classroom decisions requires more than good intentions or previous experience. Good decisions, such as what to teach, how to teach it, and how to evaluate students' achievement, are based on high-quality information. Successful teachers obtain information about their students from high-quality assessments.

And assessment involves more than testing and grading students. Assessment involves gathering and using information to improve your teaching and your students' learning. Whether you use teacher-made assessment procedures, assessments from your district's curriculum materials, or state and standardized assessments, you need to be able to explain the results correctly to students, parents, other teachers, and school administrators. Further, as you develop professionally, you may have the opportunity to participate in local and state committees concerned with assessment issues. The mainstream media as well as educational publications emphasize assessment as a major concern and consider it a newsworthy issue. It is likely to remain so for much of your professional career. This book discusses a variety of educational decisions that depend on assessments, especially in the classroom.

DISTINCTIONS AMONG ASSESSMENTS, TESTS, MEASUREMENTS, AND EVALUATIONS

The general public often uses the terms *assessment*, *test*, *measurement*, and *evaluation* interchangeably, but it is important for you to distinguish among

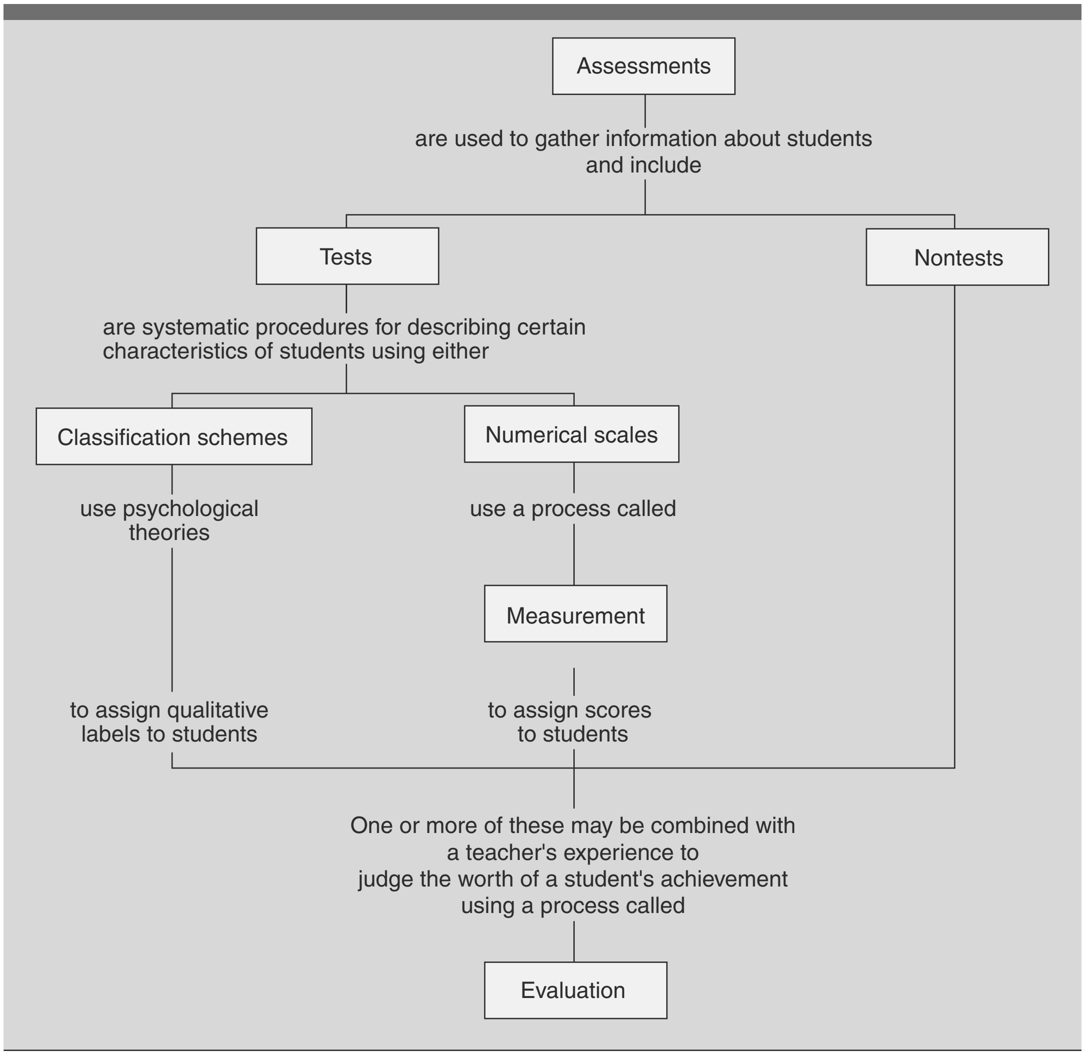

Assessments
are used to gather information about students and include
Tests
Nontests
are systematic procedures for describing certain characteristics of students using either
Classification schemes
Numerical scales
use psychological theories
use a process called
Measurement
to assign qualitative labels to students
to assign scores to students
One or more of these may be combined with a teacher's experience to judge the worth of a student's achievement using a process called
Evaluation

FIGURE 1.1 **Relationship among the terms *assessments*, *tests*, *measurement*, and *evaluation*.**

them. The meanings of the terms, as applied to situations in schools, are explained in the following paragraphs. This section explains the relationship among these terms (shown in Figure 1.1) and the way assessments inform educational decisions (Figure 1.2).

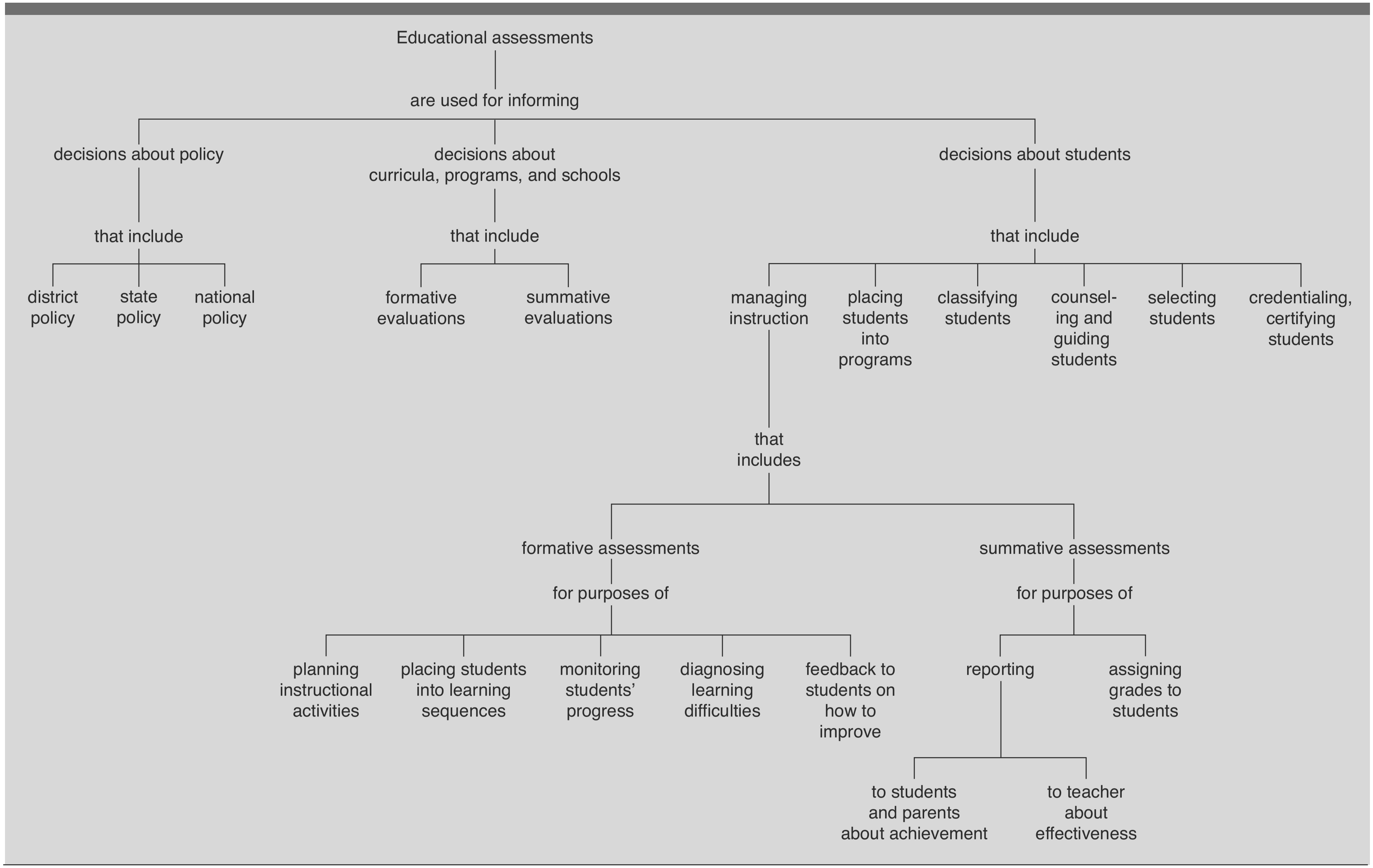
Assessments

Assessment is a broad term defined as a process for obtaining information that is used for making decisions about students; curricula, programs, and schools; and educational policy. When we say we are "assessing a student's competence," for example, we mean we are collecting information to help us decide the degree to which the student has achieved the learning targets. A large number of assessment techniques may be used to collect this information. These include formal and informal observations of a student; paper-and-pencil tests; a student's performance on homework, lab work, research papers, projects, and during oral questioning; and analyses of a student's records. This book will help you decide which of these techniques are best for your particular teaching situations.

Guidelines for Selecting and Using Classroom Assessments

Assessment is a process for obtaining information for making a particular educational decision. You should focus your assessment activities on the information you need to make particular educational decisions. This means that you need to become competent in selecting and using assessments. Here is a set of five guiding principles that

FIGURE 1.2 **Examples of types of educational decisions for which assessments may be used.**


Educational assessments
are used for informing
decisions about policy
decisions about curricula, programs, and schools
decisions about students
that include
that include
that include
district policy
state policy
national policy
formative evaluations
summative evaluations
managing instruction
placing students into programs
classifying students
counsel-ing and guiding students
selecting students
credentialing, certifying students
that includes
formative assessments
summative assessments
for purposes of
for purposes of
planning instructional activities
placing students into learning sequences
monitoring students' progress
diagnosing learning difficulties
feedback to students on how to improve
reporting
assigning grades to students
to students and parents about achievement
to teacher about effectiveness

you should follow to select and use educational assessments meaningfully.

1. Be clear about the learning targets you want to assess.
2. Be sure that the assessment techniques you select match each learning target.
3. Be sure that the selected assessment techniques serve the needs of the learners.
4. Whenever possible, be sure to use multiple indicators of achievement for each learning target.
5. Be sure that when you interpret—or help students interpret—the results of assessments, you take the limitations of such results into account.

1. *Be clear about the learning targets you want to assess.* Before you can assess a student, you must know the kind(s) of student knowledge, skill(s), and performance(s) about which you need information. The knowledge, skills, and performances you want students to learn are sometimes called learning targets or standards. The more clearly you are able to specify the learning targets, the better you will be able to select the appropriate assessment techniques.

2. *Be sure that the assessment techniques you select match the learning target.* "Do we want to evaluate students' problem posing and problem solving in mathematics? Experimental research in science? Speaking, listening, and facilitating a discussion? Doing document-based historical inquiry? Thoroughly revising a piece of imaginative writing until it 'works' for the reader? Then let our assessment(s) be built out of such exemplary intellectual challenges" (Wiggins, 1990, p. 1). The assessment techniques selected should be as practical and efficient to use as possible, but practicality and efficiency should not be the overriding considerations.

3. *Be sure that the selected assessment techniques serve the needs of the learners.* Proper assessment tools are concrete examples for students of what they are expected to do with their learning. Assessment techniques should provide learners with opportunities for determining specifically what they have achieved and specifically what they must do to improve their performance. Therefore, you should select assessment methods that allow you to provide meaningful feedback to the learners. You should be able to tell students how closely they have approximated the learning targets. Good assessment is good instruction.

4. *Whenever possible, be sure to use multiple indicators of performance for each learning target.* One format of assessment (such as short-answer questions or matching exercises) provides an incomplete picture of what a student has learned. Because one assessment format tends to emphasize only one aspect of a complex learning target, it typically underrepresents that learning target. Getting information about a student's achievement from several assessment modalities usually enhances the validity of your assessments. Matching exercises, for example, emphasize recall and recognition of factual information; essay questions emphasize organizing ideas and demonstrating writing skill under the pressure of time limits; and a monthlong project emphasizes freely using resources and research to more thoroughly analyze the topic. All three of these assessment techniques may be needed to ascertain the extent to which a student has achieved a given learning target.

5. *Be sure that when you interpret—or help students interpret—the results of assessments, you take the limitations of such results into account.* Although Guiding Principle 2 calls for increasing the authenticity or meaningfulness of the assessment techniques, assessments in schools cannot completely reproduce those things we want students to learn in "real life." The information we obtain, even when we use several different types of assessments, is only a sample of a student's attainment of a learning target. Because of this, information from assessment contains sampling error. Also, factors such as a student's physical and emotional conditions further limit the extent to which we can obtain truly accurate information. Teachers, students, and others must make decisions nevertheless. The decisions must keep an assessment's limitations in mind.

Tests

A **test** is defined as an instrument or systematic procedure for observing and describing one or more characteristics of a student using either a numerical scale or a classification scheme. *Test* is a concept narrower than *assessment*. In schools, we usually think of a test as a paper-and-pencil instrument with a series of questions that students must answer. Teachers usually score these tests by adding together the "points" a student earned on each question. By using tests this way, teachers describe the student using a numerical scale. Similarly, a preschool child's cognitive development could be observed by using the *Wechsler Preschool and Primary Scale of Intelligence* (see Chapter 18) and described as having a percentile rank of 50 (see Chapter 16). Not all tests use numerical scales. Others use systematic observation procedures to place students into categories.

Although it is natural to assume that tests are designed to provide information about an individual, this is not always true. States have testing programs designed to determine whether their *schools* have attained certain goals or standards. A federal law—the No Child Left Behind (NCLB) Act of 2001—mandates that each state use tests to evaluate whether schools are making adequate progress in improving students' achievement of the state's educational standards. Although these tests are administered to individual students, a state uses the results to measure the effectiveness of a school. In such cases, individual names are not associated with scores when reporting to the government. The "score" for the school system (or for a specific school at a specific grade level) is usually the percentage of the school's students who meet or exceed that state's standards.

Another example of an assessment program designed to survey the educational system rather than individual students is the National Assessment of Educational Progress (NAEP) (http://nces.ed.gov/nationsreportcard). The NAEP assesses the impact of the nation's educational efforts by describing what students are able to do. Assessment tasks are assigned to students on a random sampling basis, so that not every student has the same or even comparable tasks. Thus, it is not meaningful to use the scores with individual students. The assessment is intended to pool the results from all students in the sample to show the progress of education in the entire country.

The NAEP surveys are efficient ways to gather information about the average performance of a group of students because they assess each student using very few tasks, but pool the results to estimate the average. However, this gain in efficiency of assessing the group comes at the expense of not being able to describe validly the achievement of individual students.

Measurement

Measurement is defined as a procedure for assigning numbers (usually called scores) to a specified attribute or characteristic of a person in such a way that the numbers describe the degree to which the person possesses the attribute. An important feature of the number-assigning procedure in measurement is that the resulting scores maintain the order that exists in the real world among the people being measured. At the minimum this would mean, for example, that if you are a better speller than we are, a test that measures our spelling abilities should result in your score (your measurement) being higher than ours.

For many of the characteristics measured in education and psychology, the number-assigning procedure is to count the correct answers or to sum points earned on a test. Alternately, we may use a scale to rate quality of a student's product (for example, an essay or a response to an open-ended mathematics task) or performance (how well the student carries out chemistry lab procedures). (See Chapter 12 for examples.) Most measurement specialists would probably agree that although a counting or rating procedure is crude, as a practical matter, scores from assessments are useful when they are validated by using data from research (Kane, 2006).

Thus an assessment may or may not provide measurements. If a procedure describes a student by qualitative labels or categories but not by numbers, the student is assessed, but not measured in the sense used here. *Assessment* is a broader term than *test* or *measurement* because not all types of assessments yield measurements.

Evaluation

Evaluation is defined as the process of making a value judgment about the worth of a student's product or performance. For example, you may judge a student's writing as exceptionally good for his grade placement. This evaluation may lead you to encourage the student to enter a national essay competition. To make this evaluation, you would first have to assess his writing ability. You may gather information by reviewing the student's journal, comparing his writing to that of other students and to known quality standards of writing, and so on. Such assessments provide information you may use to judge the quality or worth of the student's writing. Your judgment that the student's writing is of high quality would lead you to decide to encourage him to enter the competition. Evaluations are the bases for decisions about what course of action to follow.

Evaluation may or may not be based on measurements or test results. Among others, evaluations may be based on counting things, using checklists, or using rating scales. Clearly, evaluation does occur in the absence of tests, measurements, and other objective information. You can—and probably often do—evaluate students on the basis of assessments such as systematic observation and qualitative description, without measuring them. Even if objective information is available and used, evaluators must integrate it into their own experiences to come to decisions. So degrees of subjectivity, inconsistency, and bias influence all evaluations. Testing and measurement, because they are more formal, standardized, and objective than other assessment techniques, reduce some of the inconsistency and subjectivity that influence evaluation. To say, however, that using tests and measurements (or, in general, quantitative information) "greatly improves" evaluation is itself a bias toward the technological.

Evaluation of Schools, Programs, or Materials

Not all evaluations are of individual students. You also can evaluate a textbook, a set of instructional materials, an instructional procedure, a curriculum, an educational program, or a school. Each of these things may be evaluated during development as well as after they are completely developed. The terms *formative* and *summative* evaluation are also used to distinguish the roles of evaluation during these two periods (Cronbach, 1963; Scriven, 1967). Historically, these terms arose first in the context of evaluation of schools or programs and were then applied to students. The convention has become that "formative and summative evaluation" refers to schools, programs, or materials, and "formative and summative assessment" refers to students. We will follow that convention.

Formative evaluation of schools, programs or materials is judgment about quality or worth made during the design or development of instructional materials, instructional procedures, curricula, or educational programs. The evaluator uses these judgments to modify, form, or otherwise improve the school, program, or educational material. A teacher also engages in formative evaluation when revising lessons or learning materials based on information obtained from their previous use.

Summative evaluation of schools, programs, or materials is judgment about the quality or worth of schools, or already-completed instructional materials, instructional procedures, curricula, or educational programs. Such evaluations tend to summarize strengths and weaknesses; they describe the extent to which a properly implemented program or procedure has attained its stated goals and objectives. The results of summative evaluations, more than formative evaluations, suggest whether a particular educational product "works" and under which conditions or under what degree of implementation. Summative evaluations usually are directed less toward providing suggestions for improvement than are formative evaluations.

Evaluation of Students

You may evaluate students for formative or summative purposes. **Formative assessment of students' achievement** means we are judging the quality of a student's achievement while the student is still in the process of learning. We make formative assessments of students so we can guide their next learning steps. When you ask questions in class to see whether students understand the lesson, for example, you are obtaining information to formatively evaluate their learning. You can then adjust your lesson if students do not understand. Students participate in formative assessment as well, interpreting information about their own performances to adjust their learning strategies (Popham, 2008).

Summative assessment of students' achievement means judging the quality or worth of a student's achievement after the instructional process is completed. Giving letter grades on report cards is one example of reporting your summative evaluation of a student's achievement.

HIGH-STAKES ASSESSMENT AND ACCOUNTABILITY

It may not come as a surprise to you that what you teach and how you teach it are not entirely under your control. Legally mandated external assessment programs place constraints on your teaching. You need to be aware of these as you plan your classroom instruction.

High-Stakes Testing

High-stakes assessments (tests) are used for decisions that result in serious consequences for school administrators, teachers, or students. Here are some examples:

Examples

High-Stakes Testing

Example 1. In a certain country, at the end of their secondary schooling, students must pass an examination for each subject they studied. The examinations cover the concepts and skills that are in the curriculum. Students are marked as A, B, C, D, and F for each examination. Students must get no Fs in order to be awarded the secondary school certificate. Persons without a secondary school certificate find it difficult to get a job in the country because employers see the certificate as indicating that candidates for a job have necessary minimum competencies. Students who fail may study on their own time and take the examination again, but they cannot repeat the schooling because there are only a limited number of places in secondary schools. Students must have As and Bs but no Ds or Fs to be considered for a place in one of the few universities.

Example 2. In a certain state, students must pass tests in English, writing, and mathematics before Grade 12; otherwise they cannot receive a high school diploma. They begin taking the test in Grade 10, and they may repeat the tests they failed once each year up to Grade 12. Students who do not pass all of the tests by the end of Grade 12 receive only an attendance certificate.

Example 3. In another state, students take annual state-mandated tests in reading and mathematics from Grades 3 through 11. Students do not have to pass the tests, but each school is evaluated by how well its students do. If a school's students do not show a pattern of continued improvement on the tests, the state sanctions the school by dismissing the administrative staff and perhaps some of the teachers. It turns over the running of the school to a state-appointed team until the test scores show regular improvement.

In Example 1, the consequences of assessment are quite serious for individual students: If they fail to pass all subjects they may not get a job because employers require a secondary school certificate; if they fail to do well on the examinations, they have no opportunity for attending a university. The stakes are high in Example 2, but not quite as high as in Example 1. Students can stay in school for several years, prepare for the tests, and retake the tests each year. In Example 3, there are high stakes for school administrators and teachers, but not for individual students. In fact, the tests may be low stakes for the students because there appear to be no consequences to them for doing poorly on the tests.

Accountability Testing

Although the use of high-stakes testing in the United States can be traced back to Horace Mann in the 1850s, modern high-stakes testing in the United States grew out of school reform movements that developed during the 1980s. Educational reformers and state legislators wanted to ensure that virtually all students could meet educational standards set by the state and demanded by employers. Employers needed to increase productivity and to be competitive in world markets. They needed a better-educated workforce to handle the demands of the rapidly increasing technology and greater intellectual skills needed in the workplace. State legislators considered testing to be one way of holding schools accountable for students learning the educational standards set by a state.

Assessment that is used to hold individual students or school officials responsible for ensuring that students meet state standards is called **accountability testing**. Usually accountability testing is accompanied by high-stakes consequences. A state's accountability testing may take several forms, as is shown by the examples above. A state may require both individual and school accountability, too. Check your state's education department Website for its current regulations regarding individual and school accountability.

No Child Left Behind Act

The No Child Left Behind Act is important to our discussion of high-stakes assessment because it requires states to establish challenging content standards and performance standards (referred to as *achievement standards* in the NCLB Act literature), and to demonstrate by way of tests and other assessments how well students have attained high levels of achievement on these standards. A state's failure to provide this demonstration results in loss of federal education funds that are authorized under the NCLB Act. Assessment under the NCLB Act is a school-level accountability tool.

Standards-Based Proficiency Requirements

Content standards describe the subject-matter facts, concepts, principles, and so on that students are expected to learn. **Performance standards** describe the things students can perform or do once the content standards are learned. (We discuss state standards and how to align your learning targets to them in Chapter 2.) When students are assessed on a state's standards, they are classified into one of three categories for purposes of reporting to the federal government: basic, proficient, and advanced. A state may have more than three categories, but all must be aligned to these three. Under the NCLB Act, the goal is for 100% of the students in each school to reach the proficient level or higher on the state's content and performance standards by 2014. In addition, schools must show *adequate yearly progress (AYP)* toward this goal, otherwise sanctions will be imposed.

Disaggregation An important provision of the NCLB Act is that a state must report test summaries at the school level and must disaggregate the data. **Disaggregation of test results** means that the test results for the total population of students are separated in order to report on individual subgroups of students—such as students who are poor, who are members of minority groups, who have limited English proficiency, and who have disabilities—in addition to reporting on the total student population. The reason for this requirement is that the federal government wants to ensure that states are accountable for all students learning the challenging state standards, including those in these subgroups. In some instances in the past, states reported only on the whole population of their students, thus masking the fact that some subgroups of students were not receiving quality education and were failing to meet the standards.

Assessment of Students With Disabilities Under the NCLB Act all students must be assessed, including students with disabilities and students with limited English proficiency. Ninety-five percent of students with disabilities must participate in the assessment. Students' disabilities may be used as a basis for accommodations to the assessment process when they are unable to participate under the standardized conditions set for the general student population. Further, alternative assessment methods must be found to assess those students who cannot participate even with accommodations. States are now granted some limited flexibility in adjusting content and performance standards for students with severe cognitive impairments (U.S. Department of Education, 2005).

High-Stakes Sanctions The sanctions and corrective actions that follow failure to make adequate yearly progress after 2 years include the following: (1) parents may choose to have their children attend another school in the district that is making AYP, (2) the school staff may be replaced, (3) a new curriculum may be implemented, (4) the authority of the administrative staff of the school may be changed, (5) the school year may be extended, (6) the school may be reorganized, (7) the school may be reopened as a charter school, (8) the state may contract with a private company to run the school, and (9) the school may be taken over by the state. We have already mentioned withholding federal education funding for noncompliance with NCLB. From these sanctions, you can see how high the stakes are for testing under NCLB.

Effectiveness of NCLB Whether NCLB assessment and accountability requirements have improved or hindered education is controversial. Advocates of strong accountability testing support the federal government's position that "No Child Left Behind is designed to change the culture of America's schools by closing the achievement gap, offering more flexibility, giving parents more options, and teaching students based on what works" (U.S. Department of Education, n.d.). Proponents view assessment as an objective way to ensure that all students demonstrate that learning has occurred. Some have prepared "instructionally supportive accountability" assessments that "(1) measure students' mastery of only a modest number of extraordinarily significant curricular aims, (2) describe the nature of those aims for teachers with great clarity, and (3) provide aggregatable reports of each student's attainment of every curricular aim assessed" (Popham, 2005).

Critics point to the inevitable corruption of test scores when stakes are high, including the narrowing of the curriculum to easily tested objectives whenever the focus of the school is on improving scores on tests (e.g., Nichols & Berliner, 2008). (We discuss appropriate and ethical test preparation strategies that teachers and school administrators should use in both Chapters 5 and 13 of this book.) Others point out that "for special education students and the schools that serve them, the requirements of two federal education laws and their implementing regulations, the Individuals with Disabilities Education Act (IDEA) and the No Child Left Behind Act (NCLB), are in conflict" (Phillips, 2005). A number of states criticize NCLB because adequate funds for paying for the assessments and for educational improvements were promised but never delivered in sufficient amounts (Committee on Education and the Workforce, 2005).

Teachers Coping With NCLB and Other High-Stakes Testing

In this book, we focus on using assessment results to improve your teaching and students' learning. High-stakes testing will require you to carefully determine how the content and learning targets of your teaching and your student assessments are aligned with your state's standards. Some school districts have already prepared curricular materials and sample assessments that show this alignment. Be thankful to the colleagues that preceded you! If such guidelines are available from your school office, use them as you develop your teaching and assessment plan.

We will help, too. In Chapter 2, we discuss how you can develop specific learning targets from statements of content and performance standards. In Chapter 5 we discuss some of your professional responsibilities and ethical behaviors that relate to preparing students for tests, including high-stakes accountability testing. In Chapter 6, we illustrate how to plan for assessments that are aligned with state standards and your own teaching. In Chapter 7, we discuss diagnostic and formative assessment strategies to use before and during teaching. In Chapters 8–12 we provide the knowledge you need to develop the appropriate assessments. We expand on preparing students for taking tests in Chapter 13, where we offer some practical suggestions.

The reality is that high-stakes assessment will have an impact on what and how you teach. The assessment skills you learn through your course instructor and in this book will help you with teaching and assessing in your professional practice.

ASSESSMENT AND EDUCATIONAL DECISIONS ABOUT STUDENTS

You now know that assessment provides information for decisions about students; schools, curricula, and programs; and educational policy. This section discusses several types of educational decisions made about students. It puts assessment into a broader context to give you a better idea of the purposes for which assessments are used (see Figure 1.2).

Understanding the features of different types of decisions will help you evaluate various assessment techniques that you may be thinking about using. There is no simple answer to the question, "Is this a good assessment procedure?" An assessment procedure may serve some types of decisions very well, others not so well. Understanding the different types of decisions discussed in this section will also help you explain to parents why you used various assessments with their children. Finally, although you may not be required to make all of these types of student decisions yourself, by the time your students have completed their education they will have experienced virtually all of them.

Instructional Decisions

Teaching and learning require you to constantly gather information and make decisions. Teachers make decisions about students at the rate of one every 2 to 3 minutes (Shavelson & Stern, 1981). That's about 20 decisions every class period! Sound teaching decisions require sound information. Sound assessment procedures gather sound information. Researchers estimate that teachers may spend from one third to one half of their time in assessment-related activities (Stiggins, Conklin, & Associates, 1992).

To help you think about the many decisions a teacher must make, we have organized a set of questions teachers must answer before, during, and after teaching. Examples of assessment methods that may give you useful information for making the decisions are listed in parentheses after each question.

Decisions Before Beginning Teaching

1. What content do I need to cover during the next day, week, month, marking period, and so on? (Possible assessment methods: *Review the curriculum, the syllabus, and the textbook; examine copies of the standardized tests my students will need to pass.*)

2. What abilities (cultural background factors, interests, skills, etc.) of my students do I need to take into account as I plan my teaching activities? (Possible assessment methods: *Informal observation of students during class discussions; conversations with students and students' previous teachers; studying*

students' permanent records to see their scholastic aptitude test results, past grades, and standardized test results; my knowledge of the student's personal family circumstances.)

3. What materials are appropriate for me to use with this group of students? (Possible assessment methods: *Class discussions in which I observe students' motivations, interests, beliefs, and experience with the topics I will teach, and their attitudes toward learning the topics; results from short pretests I administer; my study of the students' permanent records to learn the previous teacher's evaluations and the students' standardized achievement test results*.)

4. With what learning activities will my students and I need to be engaged as I teach the lesson (unit, course)? (Possible assessment methods: *My review of the types of activities I used previously that stimulated the interests of students; my knowledge of typical student learning progressions in this area; my analysis of the sequence of the learning activities students will follow; my review of how well the students achieved when those activities were used previously*.)

5. What learning targets do I want my students to achieve as a result of my teaching? (Possible assessment methods: *My review of statements of goals and learning objectives; my review of test questions students should be able to answer; my review of the things students should be able to do and of the thinking skills students should be able to demonstrate after learning*.)

6. How should I organize and arrange the students in the class for the upcoming lessons and activities? (Possible assessment methods: *My informal observation of students with special learning and social needs; my recollection of students' behavior during previous learning activities; information about what classroom arrangements worked best in the past when my students were learning similar targets*.)

Decisions During Teaching

1. Is my lesson going well? Are students catching on (i.e., learning)? (Possible assessment methods: *My observations of students during learning activities; student responses to questions I have asked them; my observations of students' interactions*.)

2. What should I do to make this lesson (activity) work better? (Possible assessment methods: *My diagnosis of the types of errors students made or misconceptions students have; searching my memory for alternative ways to teach the material; my identifying which students are not participating or are acting inappropriately*.)

3. What feedback should I give each student about how well he or she is learning? (Possible assessment methods: *My informal observation and experience on the amount and type of feedback information different students require; information about how close each student has come to achieving the learning target; students' homework and quiz results; my interviews of students*.)

4. Are my students ready to move to the next activity in the learning sequence? (Possible assessment methods: *My informal observation and checking of students' completed work and questioning students about their understanding; my analysis of students' homework, quizzes, and test results; results of student self-assessment*.)

Decisions After a Teaching Segment

1. How well are my students achieving the short- and long-term instructional targets? (Possible assessment methods: *My classroom tests, projects, observations, interviews with students; my analysis of standardized test results*.)

2. What strengths and weaknesses will I report to each student and to his or her guardian or parent? (Possible assessment methods: *My observations of each student's classroom participation; my review of each student's homework results; my review of each student's standardized achievement and scholastic aptitude test results when they become available; my review of information about a student's personal family circumstances*.)

3. What grade should I give each student for the lesson or unit, marking period, or course? (Possible assessment methods: *My combining results from classroom learning activities, quizzes, tests, class projects, papers, labs, etc.; my observation about how well the student has attained the intended learning targets*.)

4. How effectively did I teach this material to the students? (Possible assessment methods: *My review of summaries of the class's performance on the important instructional targets and on selected questions on standardized tests, and of how well the students liked the activities and lesson materials*.)

5. How effective are the curriculum and materials I used? (Possible assessment methods: *My review of summaries of informal observations of students'*

interests and reactions to the learning activities and materials; of the class's achievement on classroom tests that match the curriculum; and of several past classes' performance on selected areas of standardized tests.)

These lists of questions and assessments are not exhaustive; you may wish to list others. These examples illustrate that your teaching decisions require you to use many different types of information. Further, they illustrate that the exact type of information you need varies greatly from one teaching situation to the next.

Instructional Diagnosis and Remediation Sometimes the instruction an individual student receives is not effective: The student may need special remedial help or special instruction that relies on alternative methods or materials. Assessments that provide some of the information needed to make this type of decision are called **diagnostic assessments**. Diagnostic decisions center on the question, "What learning activities should I use to best adapt to this student's individual requirements and thereby maximize the student's opportunities to attain the chosen learning target?" Diagnosis implies identifying both the appropriate content and the types of learning activities that will help a student attain the learning target (Glaser & Nitko, 1971; Nitko, 1989; Nitko & Hsu, 1974).

Feedback to Students Assessments also provide feedback to students about their learning. Feedback, however, is likely to improve learning only under certain conditions. Simply assessing students and reporting the results to them is not likely to affect their performance. Learners must review both correct and incorrect performance and, in addition, be able to correct their incorrect performance. In other words, feedback must give specific guidance to students about what they must do to improve their learning. Therefore, teachers who give students only their grade on a paper or test are not providing enough feedback to help students improve.

Assessments can be used to provide feedback that helps learning, provided you integrate them into your instructional process. Feedback from classroom assessment procedures will not help your students learn if the students lack a command of the prerequisite learning and/or have comprehended little or nothing of the lesson prior to the assessment. It is especially important that you correct students' errors—or that the students correct their own errors—before going on to new instruction. Similarly, frequent feedback during the lesson is essential. Additional discussion of feedback appears in Chapter 7.

Feedback to the Teacher Remember that assessments provide feedback to the teacher about how well students have learned and how well the teacher has taught. Of course, if students have failed to grasp important points, the teacher should reteach the material before proceeding to new material.

Modeling Learning Targets Assessments serve as examples for students by showing them what you want them to learn. Assessments, as well as other assignments, should therefore *embody* the learning target (Shepard, 2006) so that students get an accurate and clear idea of what they are to learn. Students can compare their current performance on the learning target with the desired performance. You may teach them to identify the way(s) in which their current performance matches the expected performance and how it is deficient. Your teaching can focus on how to remedy the deficiencies. In these ways, good assessment is good instruction. Also, as students evaluate their own performance, you may teach them the appropriate criteria for judging how well they are learning as well as teaching them what is important to learn.

Motivating Students Assessments may also motivate students to study. Unfortunately, some teachers use this form of accountability as a weapon rather than as a constructive force. Teachers may hope that using an assessment as a possible threat will encourage their students to take studying seriously. Sometimes teachers use the "surprise quiz" or "pop quiz" in this manner to encourage more frequent studying and less cramming.

Studies have not justified use of assessments this way. Rather, assessments ought to be viewed in a more positive light: as tools for instruction and feedback to students (Glaser & Nitko, 1971). Also, teachers or parents who stress test performance as the sole or major criterion for school success may create undue test anxiety for students. As a result, students may perform less well in the long run.

Assigning Grades to Students One of the most obvious reasons for giving classroom assessments is to help you assign grades to students. Although teachers continually assess their students' progress in informal ways, they also must officially record their evaluations of students' progress through grades. The grades or symbols (A, B, C, etc.) that you report represent your summative evaluations or judgments about how well your students have achieved important learning targets. Use a mixture of assessment formats to provide the information you need to make these evaluations. Good teaching practice and common sense indicate that grades should be based on more than test scores. Many teachers, however, fall back on test scores alone to justify the grades they assign. Assigning grades involves evaluative decisions, and judgments are often difficult to justify and explain. Tests, especially those of the objective variety, seem to reduce judgment and subjectivity, even though this is not necessarily true. A more complete discussion of grading, including suggestions for assigning grades, appears in Chapter 14.

Selection Decisions

Most people are familiar with **selection decisions**: An institution or organization decides that some persons are acceptable, whereas others are not; those who are unacceptable are rejected and are no longer the concern of the institution or organization. This feature—rejection and the elimination of those rejected from immediate institutional concern—is central to a selection decision.

An educational institution often uses test scores as one component for selection decisions. For example, college admissions are often selection decisions: Some candidates are admitted and others are not; those who are rejected are no longer the college's concern. Some critics may argue, however, that those rejected should still be of concern to society generally.

When an institution uses an assessment procedure for selection, it is important to show that candidates' results on these assessments bear a significant relationship to success in the program or job for which the institution is selecting persons. If data do not show that these assessment results can distinguish effectively between those candidates likely to succeed and those unlikely to succeed, then these assessment procedures should be improved or eliminated. In fact, it may be illegal to continue to use assessment results that bear no relationship to success on the job (Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, Department of Labor, & Department of the Treasury, 1979; United States Supreme Court, 1971.).

Selection decisions need not be perfect to be useful, however. Assessment results cannot be expected to have perfect validity for selection, or any other, decisions (see Chapter 3). Figure 1.3 illustrates the use of imperfect assessments in selection. Some applicants would have been successful had

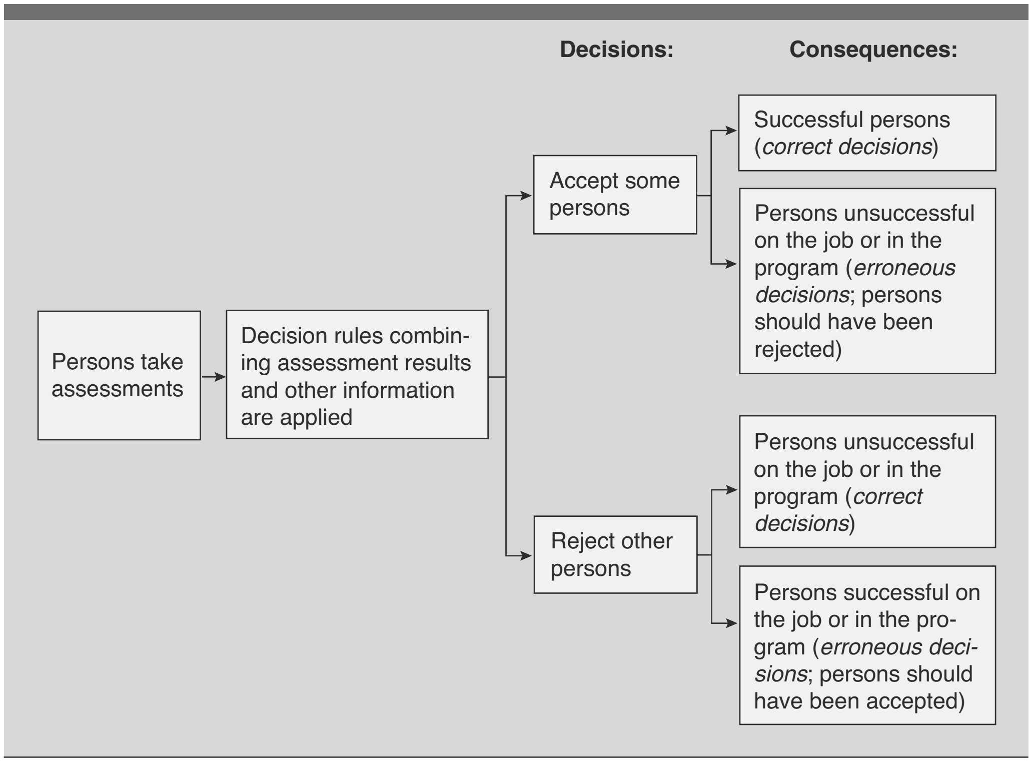

Decisions:
Consequences:
Persons take assessments
Decision rules combining assessment results and other information are applied
Accept some persons
Reject other persons
Successful persons (*correct decisions*)
Persons unsuccessful on the job or in the program (*erroneous decisions*; persons should have been rejected)
Persons unsuccessful on the job or in the program (*correct decisions*)
Persons successful on the job or in the program (*erroneous decisions*; persons should have been accepted)

FIGURE 1.3 A simplified illustration of how a selection situation uses assessments and the consequences of those decisions. The assessments and the decision rules are evaluated in terms of their consequences.

they been selected instead of rejected (false negative decisions); and some, even though they were accepted, turned out to be unsuccessful (false positive decisions). Assessments can be evaluated, then, in terms of the consequences of the decisions made when using them. This subject is taken up in Chapter 3.

Placement Decisions

Placement decisions are characterized as follows: Persons are assigned to different levels of the same general type of instruction, education, or work; no one is rejected, but all remain within the institution to be assigned to some level. Students not enrolled in honors sections, for example, must be placed at other educational levels. Or, first-grade students with low scores on a reading readiness test cannot be sent home. They must be placed in appropriate educational levels and taught to read. You may recognize a decision as a placement decision by noting whether the institution must account for all candidates. The rejection of candidates and their elimination from the institution's concern that occurs with selection decisions is not possible in placement decisions.

Many, if not most, decisions in schools are placement decisions. Educators who use the language of selection often are using the language incorrectly. On closer examination, they are speaking about placement decisions. For example, when an educator speaks of "screening" students for a gifted and talented program, the decisions are actually placement decisions because their ultimate purpose is to place all students in appropriate educational programs. The schools are not free to teach some students and to reject the rest. If one instructional method is inappropriate for a particular student, then an appropriate alternative method needs to be found. In the end, all students are taught, and must learn.

Classification Decisions

Sometimes we must make a decision that results in a person being assigned to one of several different but unordered categories, jobs, or programs. These types of decisions are called **classification decisions**. For example, educational legislation concerning persons with disabilities has given a legal status to many labels for classifying children with disabilities and strongly encourages classifying them into one (or more) of a few designated categories. These categories are unordered (blindness is not higher or lower than deafness), so these are classification decisions rather than placement decisions.

You may consider *classification* as a more general term that subsumes *selection* and *placement* as special cases. *Classification* refers to cases in which the categories are essentially unordered, *placement* refers to cases in which the categories represent ordered levels of education without rejection, and *selection* refers to cases in which students are accepted or rejected. This book considers the three types of decisions separately.

Counseling and Guidance Decisions

Assessment results frequently assist students in exploring, choosing, and preparing for careers. A single assessment result is not used for making guidance and counseling decisions. Rather, a *series of assessments* is administered, including an interview, an interest inventory, various aptitude tests, a personality questionnaire, and an achievement battery. Information from these assessments, along with additional background information, is discussed with the student during a series of counseling sessions. This facilitates a student's decision-making process and provides a beginning for exploring different careers. Exploring career options is likely to involve an ongoing and changing series of decisions that occur throughout a person's life.

Credentialing and Certification Decisions

Credentialing and certification decisions reflect whether a student has attained certain standards of learning. Student certification decisions may focus on whether a student has attained minimum competence or obtained a high standard, depending on the legal mandate. Certification and credentialing may be mandated by a state's legislation or may be voluntary. If a state law mandates that students achieve certain standards of performance, most often students are administered an assessment procedure created at the state level. Those who meet the standards are awarded a credential (such as a high school diploma).

These certification assessment procedures present special problems for validation. Individual students cannot reasonably be held accountable for instruction that the teacher failed to deliver or which was delivered poorly, even though, on the average, teaching was adequate. A critical point, therefore, is whether the quality of instruction corresponds to what the assessment procedure covers. The closer the correspondence, the fairer the certification is to the student. If students did not have the opportunity to learn how to perform the tasks that appear on the certification assessment procedure, either because a specific school lacked the necessary resources or a particular teacher failed to deliver appropriate instruction, the assessment-based certification process seems inherently unjust.

Often, it is not easy to resolve conflicts about what has been taught and what should appear on an assessment procedure. For example, suppose a state holds students accountable for a reading list of "important works." Suppose one group's teacher did not explain these works directly, but another group's teacher did. Should the first group be held accountable? Further, high standards in some states may require students to apply knowledge to new situations, to solve new problems, and to exhibit creativity. To assess application to new situations, assessments must include tasks that are unfamiliar or novel for students. This is accomplished by deliberately making the assessment materials different from the materials used during teaching (otherwise they would not be "novel"). The validity question is, "Is this fair to students?"

Another example is using assessments for teacher certification. Some states assess preservice teachers' knowledge using paper-and-pencil tests. Often a battery of tests will include basic skills, general knowledge, and professional knowledge. Some tests also include separate assessment of specialty areas such as biology, elementary education, and teaching students with hearing impairments. Some states also evaluate a teacher's classroom performance through observation or by assigning a master teacher to mentor a beginning teacher. The National Board for Professional Teaching Standards (NBPTS) has developed assessment procedures to certify experienced teachers who are outstanding in their teaching skills (see http://www.nbpts.org). Teachers are assessed in many areas and educational levels. Unlike the state-mandated assessments, NBPTS certification is voluntary. Also, the assessment procedures are heavily performance based. That is, teachers may submit portfolios of documents, student work samples, and perhaps videotapes to demonstrate their teaching competence. Their teaching may also be observed, and they may come to an assessment center to participate in simulated teaching activities such as instructional design, group interaction, parent-teacher conferences, peer collaboration, and staff development.

ACQUIRING THE KNOWLEDGE AND SKILLS TO ASSESS STUDENTS

To help you evaluate your present level of competence and focus on important areas of assessment skills, the American Federation of Teachers et al. (1990) published *Standards for Teacher Competence in Educational Assessment of Students*. These standards are somewhat dated now. Most importantly, they do not address formative assessment skills like being able to help students generate and use assessment information for their own learning. Appendix A presents our synthesis of the various knowledge and skills that, taken together, comprise what today would be called "assessment literacy" for teachers.

CONCLUSION

This chapter introduced you to basic assessment terms and concepts and basic types and purposes of educational decisions. It would not be exaggerating to say that appropriate assessment information should support everything teachers and administrators do in schools. The remainder of this book is devoted to developing the knowledge and skills you will need to accomplish that well. In Chapter 2 we turn to defining instructional goals or learning targets, which are the foundation on which formative and summative assessment, as well as instruction, must be based.

EXERCISES

1. Self-reflect on a specific lesson you have taught or would like to teach. Make a list of the decisions you made (or need to make) before, during, and after this lesson. Next to each decision, identify how you will obtain the information needed to make the

decision. What criterion might you use to judge the quality of each piece of information?

2. Decide whether each of the following statements is true or false. Defend your answers.
   a. To make evaluations, one must use measurements.
   b. To measure an important educational attribute of a student, one must use a test.
   c. To evaluate a student, one must measure that student.
   d. To test a student, one must measure that student.
   e. Any piece of information a teacher obtains about a student is an assessment.
   f. To evaluate a student, one must assess that student.
3. Check education magazines and newspapers during the past 4 months for articles on NCLB. (You may want to use their Websites to identify articles; for example go to http://www.edweek.org.) Also check the Websites of teachers' unions, state governors' organizations, organizations of state boards of education, and parent advocacy groups. Who are the persons or agencies concerned about NCLB? What are the major issues with which they are concerned? Summarize your results, present them to your class, and compare your results with those of the other students. What are the issues that teachers need to address in their classrooms that are related to NCLB?
4. Classify each of these statements as reflecting a selection, classification, placement, career guidance, diagnostic/remediation, or certification decision. Defend your answers.
   a. After students begin kindergarten, they are given a battery of perceptual skills tests to decide which children should receive special perceptual skills training and which should remain in the "regular" program.
   b. A child study team decides whether each child who has been administered a series of screening tests should be included in a particular category of disability (students with hearing impairments, learning disabilities, etc.).
   c. After a school psychologist assesses a student, local education authorities assign the student to the resource room, where the teacher for students with learning disabilities gives the student special instruction each day.
   d. Each graduate of this department of education is required to take and pass the state's test before being allowed to teach in the schools.
5. Self-reflect on each of the teacher assessment knowledge and skills found in Appendix A. Under each standard, describe the kinds of competence you now have and those that you hope to have at the end of this course.

Describing the Goals and Learning Targets of Instruction

KEY CONCEPTS

1. Learning objectives or learning targets focus instruction and assessment, and they also focus students and teachers, on the knowledge and skills intended for learning.
2. Different levels of specificity are used for statements of learning goals, state standards, content and performance standards, general and specific learning targets, and developmental and mastery learning targets.
3. Sources for locating learning targets include state standards, instructional materials, and professional associations.
4. Taxonomies of thinking skills help you get the most out of your learning targets and assessment tasks.
5. Before teaching, list and evaluate all your learning targets for a unit or course.
6. Specific learning targets should be student centered, performance centered, and content centered.
7. Align both instruction *and* assessment to your learning targets.

IMPORTANT TERMS

affective domain

analysis, application, comprehension, evaluation, knowledge, synthesis

cognitive domain

conceptual knowledge, factual knowledge, procedural knowledge, and metacognitive knowledge

content centered

developmental learning targets

educational goals

general learning target

learning objective

learning target

mastery learning targets

performance centered

psychomotor domain

specific learning target

standards

student centered

taxonomies of instructional learning targets

From Chapter 2 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

IMPORTANCE OF SPECIFYING LEARNING TARGETS

A **learning objective**, also sometimes called a **learning target**, specifies what you would like students to achieve when they have completed an instructional segment. The goal of teaching should involve more than "covering the material" and "keeping students actively engaged." The focus of your teaching should be on student achievement as well as on the learning process. So, your learning targets should state what students ought to be able to do, value, or feel after you have taught them.

Some learning targets are *cognitive,* meaning that they deal primarily with intellectual knowledge and thinking skills. For example, you may want students to read a claim made by a political figure and determine whether there is evidence available to support that claim. Other learning outcomes are *affective,* meaning that they deal with how students should feel or what they should value. For example, you may want students to value the right to vote in elections over other activities competing for their time. Yet other learning targets are *psychomotor,* meaning that they deal primarily with motor skills and physical perceptions. For example, you may want students to set up, focus, and use a microscope properly during a science investigation of pond water.

Deciding the specific targets you expect students to achieve in their learning is one important step in the teaching process. Instruction may be thought of as involving three fundamental but interrelated activities:

1. Deciding what students are to learn.
2. Carrying out the actual instruction.
3. Evaluating the learning.

Activity 1 requires you to articulate in some way what you expect students to be able to do after you have taught them. Usually, you do this by specifying learning targets or by providing several concrete examples of the tasks students should be able to do to demonstrate that the learning targets have been reached. Activity 1 informs you and the students about what is expected as a result of teaching and studying. Your understanding of the learning targets guides your teaching and provides a criterion for deciding whether students have attained the desired change.

Activity 2 is the heart of the teaching process itself. Here you provide the conditions and activities for students to learn. These include formative assessment procedures like monitoring students' progress and giving them feedback on what they need to improve their achievement of the learning targets.

Activity 3, evaluating whether learning has occurred, is summative assessment. Through it you and your students come to know how well the learning targets have been reached. The more clearly you specify the learning targets, the more directed your teaching efforts and your students' learning efforts will be.

These three fundamental activities are interactive rather than a straight one–two–three process. Setting clear learning targets helps you plan your teaching efficiently, conduct your instruction—whether whole-class, differentiated by groups, or individualized—effectively, and assess student outcomes validly. Assessing and evaluating students using clearly specified learning targets provides you with information about how to guide students' learning and how effective your instruction has been. This information, in turn, may be used to adjust your teaching, to plan the next instructional activities, or to better specify the instructional targets. Setting clear learning targets also helps you communicate them to others.

Before you can design procedures to evaluate students' learning, you should have clearly in mind the students' performances you want to evaluate. If you are not clear on which important learning outcomes you want to evaluate, it is hardly possible to make a valid assessment of those outcomes. Statements of specific learning targets are important for the following four aspects of classroom assessment:

1. *The general planning for an assessment procedure* is made easier by knowing the specific outcomes you wish students to achieve.
2. *Selecting and creating assessment procedures* depend on your knowing which specific achievements you should assess.
3. *Evaluating an existing assessment procedure* is easier when you know the specific learning targets.
4. *Properly judging the content relevance of an assessment procedure* requires you to know the specific achievements you should assess (see Chapter 3).

EDUCATIONAL GOALS, STATE STANDARDS, AND LEARNING TARGETS

This section discusses several closely related concepts. You might find it helpful to refer to Figure 2.1 when studying them.

Educational Goals Versus Specific Learning Targets

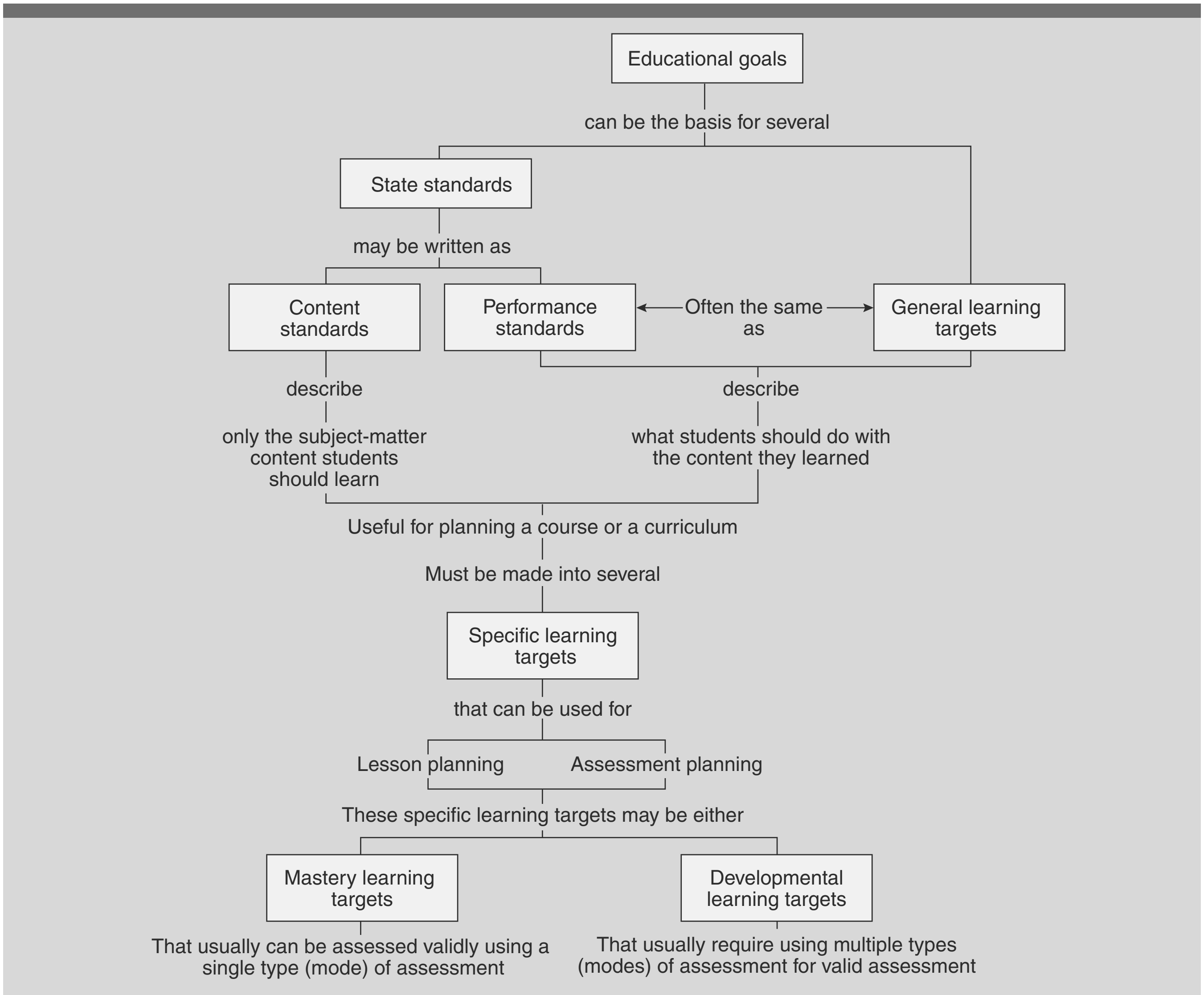
Schooling and other organized instruction help students attain educational goals. One of the many ways to define educational goals is that they "are those human activities which contribute to the functioning of a society (including the functioning of an individual in society), and which can be acquired through learning" (Gagné, Briggs, & Wager, 1988, p. 39).

Educational goals are stated in broad terms. They give direction and purpose to planning overall educational activities. Examples of statements of broad educational goals appear in reports prepared by state departments of education, local school systems, and associations such as the National Council of Teachers of Mathematics, the American Association for the Advancement of Science, and the Association of American Geographers. Here is one example of an educational goal:

Example

Every student should acquire skills in using scientific measurement.

FIGURE 2.1 **Relationships among the concepts of standards, goals, and learning targets.**


Educational goals
can be the basis for several
State standards
may be written as
Content standards
Performance standards
Often the same as
General learning targets
describe
only the subject-matter content students should learn
describe
what students should do with the content they learned
Useful for planning a course or a curriculum
Must be made into several
Specific learning targets
that can be used for
Lesson planning
Assessment planning
These specific learning targets may be either
Mastery learning targets
Developmental learning targets
That usually can be assessed validly using a single type (mode) of assessment
That usually require using multiple types (modes) of assessment for valid assessment

These types of broad goals are organized into subject-matter areas such as mathematics and history. The broad goals, and statements of subject-matter area and content-specific thinking processes, serve as a curriculum framework within which you and other educators can define specific learning targets. State education agencies take the process further by publishing expected learning outcomes or *standards,* and your school is held accountable for students' achieving these particular standards. You can obtain a copy of your state's standards from your school principal, central administration office, or your state's education department Website.

General Learning Targets Versus Specific Learning Targets

There is an appropriate level of specificity for stating learning targets. If the description of a target is stated too broadly, teachers cannot use it for developing lesson plans and assessment procedures. The previously stated educational goal, for example, may help communicate a general educational aim, but is too broadly stated to be immediately useful to plan lessons and assessments.

A general learning target is a statement of an expected learning outcome that is derived from an educational goal. General learning targets are more specific than educational goals and usually clear enough for general planning of a course. However, they need to be made more specific before they can become learning targets that you can use when planning lessons. The following example of a general learning target might be stated for a primary school science unit on measurement in the metric system:

Example

Acquire the skills needed to use common instruments to measure length, volume, and mass in metric units.

When teaching students and assessing their attainment of this general learning target you may need to break it down into two or more specific learning targets. A specific learning target is a clear statement about what students are to achieve at the end of a unit of instruction. Here are three examples of specific learning targets that are derived from the preceding general learning target:

Example

1. Measure the length of objects to the nearest tenth of a meter using a meter stick.
2. Measure the mass of objects to the nearest tenth of a kilogram using a simple beam balance and one set of weights.
3. Measure the volume of liquids to the nearest tenth of a liter using a graduated cylinder.

Danger of Overly Specific Learning Targets

When learning targets are made more specific, the achievement you are to teach and to assess becomes clear. But beware of overspecificity. Long lists of very narrow "bits" of behavior can fragment the subject to be taught. The following examples show learning targets that are too specific, along with a suggested revision:

Example

The student is able to:

Too specific: Estimate the number of beans in a jar.

Better: Solve practical problems using calculations and estimation.

Rationale: "Beans in a jar" is not the real target of learning. Rather, it is but one of the many possible tasks that a student should complete to demonstrate achievement of estimation and calculation. The learning target statement should describe this less specific achievement.

Example

The student is able to:

Too specific: Explain the meaning of the term cold front.

Better: Explain the meaning of key weather terms.

Rationale: "Cold front" is only one of several key weather terms that are included in a unit. Listing a separate learning target for each term taught in the unit fragments the unit's focus on general weather terminology.

A second danger is that lists of specific objectives may become too long and be ignored. Identify a few of the most important learning targets for each instructional unit and focus on these.

Creating Assessments That Require Students to Use Combinations of Learning Targets

It is important, too, to create learning and assessment situations that require students to use combinations of specific skills and knowledge to perform complex tasks and solve real-life problems. Figure 2.2 shows a beans-in-a-jar problem. In solving this problem students are expected to use several specific skills and knowledge (listed at the upper right of the figure) to accurately estimate the number of beans in the jar. "Beans in a jar" is not the learning target itself, of course. Rather, it is only one example of many possible tasks in which the learning target is to apply a combination of proportional reasoning, estimation, measurement, and other skills to solve complex problems.

Notice that in this example, the most important outcomes teachers should assess are the processes and strategies students use to solve these problems. The criteria for these are listed under "Criterion-referenced interpretations" in Figure 2.2. An assessment procedure that focuses exclusively on the degree of correctness of students' answers to tasks like this would be invalid because it misses assessing the processes that students use.

State Standards Versus Learning Targets

Standards

Standards are statements about what students are expected to learn. Some states call these statements *essential skills, learning expectations, learning outcomes, achievement expectations,* or other names. The NCLB Act requires all states to specify achievement standards and to assess students' attainment of them.

Often there are two sets of achievement standards. *Content standards* are statements about the subject-matter facts, concepts, principles, and so on that students are expected to learn. For example, a standard for life science might be, "Students should know that the cell nucleus is where genetic information is located in plants and animals." *Performance standards* are statements about the things students can perform or do once the content standards are learned. For example, "Students can identify the cell nucleus in microscopic slides of various plant and animal cells."

State education departments prepare standards used in schools. Local school districts are required to teach students to achieve these standards and are held accountable for students achieving them through the state's assessment system. Professional organizations can prepare standards, too. These organizations try to influence what is taught by publicly promoting their own standards. Examples of professional organizations with published standards are the National Academy of Sciences, National Council of Teachers of English, and National Council of Teachers of Mathematics. Most standards from professional organizations can be found on the organizations' Websites.

States' standards vary greatly in their quality and degree of specificity. Not all states have done a good job of writing standards. There seems to be no "standard" way to write standards.

In the past, some states have established standards for only some grade levels (e.g., 4th, 8th, 10th, and 12th grades). Under the NCLB Act states must have standards for Grades 3 through 12 in reading/language arts and science. In response to this requirement, most states have prepared many standards for each grade level. Others, however, are experimenting with writing a select few standards for each grade and focusing assessment and teaching on these (Popham, 2005).

Learning Targets

As you may have gathered from the preceding paragraphs, a state's standards are really learning targets. After officially adopting a state's standards, a school must make sure all students are taught and achieve those standards. Most state standards are written at a fairly general level. The better-written state performance standards are essentially the same as the general learning targets that we discussed earlier. You will need to break down each standard into two or more specific learning targets to teach and assess them. Thus, all of the material in this chapter applies to teaching and assessment whether your school and state use "objectives," "learning targets," or "standards." The example below shows how specific learning targets are developed from a state standard and compares statements of standards, general learning targets, and specific learning targets for third-grade reading in one school district:

FIGURE 2.2 **An assessment task used as a benchmark by the Toronto Board of Education.**

BEANS IN A JAR
Applying rate and ratio

In the task for this benchmark, students were first shown a jar filled with beans and asked to estimate the number of beans. They were then asked to work out the number of beans more accurately using any of the following materials:a calculator, a balance scale and masses, a ruler, a graduated cylinder, and a transparent centimeter-squared grid. They were told they could count some but not all of the beans. If the students did not know how to proceed, the evaluators suggested they weigh a small handful of beans. The students were asked to keep an ongoing record of their solutions. After they had solved the problem they were asked to describe the problem and their solutions.

Key objectives from the Ontario Ministry of Education and Toronto Board guidelines

- Apply ratio and rate in problem solving
- Consolidate conversions among commonly used metric units
- Collect and organize data
- Consolidate and apply operations with whole numbers and decimals with and without a calculator
- Apply estimation, rounding and reasonableness of results in calculations, in problem solving and in applications
- Develop facility in communication skills involving the use of the language and notation of mathematics
- Develop problem-solving abilities

Norm-referenced interpretations		Task score	Criterion-referenced interpretations
20% of the students scored (80% scored lower than 5)	5	5	The student understands the problem and immediately begins to search for a strategy, perhaps experimenting with different methods and materials before proceeding. The student monitors the solution as it develops and may check and remeasure. The student uses the materials efficiently and accurately and keeps a good record of the data. All the calculations are performed accurately and a reasonable answer is produced. The student gives a clear explanation of the solution demonstrating sound reasoning with proportions. The student takes ownership of the task and enjoys its challenge.
19% of the students scored (61% scored lower than 4)	4	4	The student may make some false starts and may be helped by the evaluator to get focused. The student may use some materials to no purpose or inaccurately, perhaps confusing volume and mass.The student reasons with proportions correctly. Although stuck at various points in the solution, the student perseveres and usually produces a reasonable answer. The student usually gives a clear explanation and enjoys the activity.
20% of the students attained (41% scored lower than 3; the average score is 3.0)	3	3	There is some confusion in one or more aspects of the solution to the problem. The student may confuse units, make arithmetic errors or perform incorrect operations. The student may have some idea of proportionality but is unable to use it correctly. The student does not use the materials to the best advantage. The student seeks assistance from the evaluator. Although not totally confident, the student may persevere in an attempt to arrive at an answer to the problem.
24% of the students attained (17% scored lower than 2)	2	2	The student may make a start at solving the problem but is unable to complete a solution. The student may repeatedly switch methods and materials, and be unable to find an effective strategy. There is considerable confusion with units and the interpretation of various measurements. The student usually guesses at the operations that should be performed with the data. The student lacks confidence and seeks a great deal of assistance from the evaluator.
17% of the students attained	1	1	The student may estimate the number of beans but gives no response or very limited response to working out the number more accurately.

Source: Adapted from John L. Clark (1992). The Toronto Board of Education's Benchmarks in Mathematics. *The Arithmetic Teacher: Mathematics Through the Middle Grades, 39*(6), pp. 51–55. Adapted by permission.

Example

State standard

- Communicate well in writing for a variety of purposes.

General learning target

- Write for narrative, persuasive, imaginative, and expository purposes.

Specific learning targets

- Explains the difference between narrative, persuasive, imaginative, and expository writing purposes.
- Applies prewriting skills and strategies to generate ideas, clarify purpose, and define audience before beginning to write.
- After receiving feedback on the first draft in the areas of ideas, organization, voice, word choice, and sentence fluency, uses the feedback to revise the draft.
- Reviews and revises the second draft for grammatical correctness and proper use of standard writing conventions.

Specific Learning Targets as Mastery Statements

Assessment focuses on what you can see students doing. From this observation you will infer whether they have attained the learning targets. For example, a high school biology unit on living cells may have as a general learning target that students should "learn the organizations and functions of cells." But what can the student do to demonstrate learning of this general target? There may be several answers to this question, each phrased as a specific instructional objective and each describing what a student can do, as shown in the following example:

Example

1. The student can draw models of various types of cells and label their parts.
2. The student can list the parts of a cell and describe the structures included in each.
3. The student can explain the functions that different cells perform and how these functions are related to each other.

Statements of what students can do at the end of instruction may be called mastery learning targets. They have also been called *specific learning outcomes* and *behavioral objectives*.

Mastery Learning Targets Versus Developmental Learning Targets

Some skills and abilities are more aptly stated at a somewhat higher level of abstraction than mastery learning targets to communicate that they are continuously developed throughout life. Consider the following examples:

Examples

1. Combine information and ideas from several sources to reach conclusions and solve problems.
2. Analyze and make critical judgments about the viewpoints expressed in passages.
3. Use numerical concepts and measurements to describe real-world objects.
4. Interpret statistical data found in material from a variety of disciplines.
5. Write imaginative and creative stories.
6. Use examples from materials read to support your point of view.
7. Communicate your ideas using visual media such as drawings and figures.

Because of the lifelong nature of these targets they may be called *developmental objectives* (Gronlund & Brookhart, 2009) or **developmental learning targets**.

At first glance, it might seem that all one needs to do is to insert a "can do" phrase in front of each of the preceding statements to transform them to mastery learning targets. However, it is not that simple. First, each statement represents a broad domain of loosely related (not highly correlated) performances. Second, each statement represents skills or abilities typically thought of as developing continuously to higher levels rather than the all-or-none dichotomy implied by the mastery learning targets.

The Problem of a Broad, Heterogeneous Domain

Consider Developmental Learning Target 2 in the previous list. Now, think about questions you could ask students to assess how well they have achieved this learning target. Your questions need to require students to analyze a reading passage and make inferences based on information in it. The example below shows three possible questions. These questions are passage-based items from the *National Assessment of Educational Progress* civics test. The numbers in the brackets are the percentage of 12th-grade students who answered each question correctly.

Example

1. In what way does the article show one of the strengths of federalism? [32%] (2006-12C7, question 9)
2. In what fundamental way do the two quotes above show different understandings of the rights of citizens? [51%] (2006-12C7, question 3)
3. The events at Central High School in Little Rock showed that. . . . [60%] (2006-12C5, question 17)

You can see that each question refers to a different passage with different viewpoints expressed. Further, the percentage of students answering one question is quite different from the percentage answering another. Studies of these types of questions show that those who answer one question right are not necessarily the same students who get another question right (Forsyth, 1976).

We can conclude from this that Developmental Learning Target 2 represents a broad domain of reading passages and that mastering one part of the domain does not mean mastering another. This is the case with developmental learning targets like those listed previously.

The Issue of Continuous Development of Skill

The second concern, the continuous or developmental nature of these learning targets, stems from the fact that even the simplest developmental objective is a matter of degree. Continuous development is possible throughout life. All we can reasonably expect to do for a particular course or unit of instruction is to identify a sample of specific learning outcomes that represent degrees of progress toward the objectives. The essential concern here is that the skills represented by these learning targets are complex, the number of tasks that can be used to demonstrate learning is vast, and each represents goals to work toward continuously rather than to master completely (Gronlund & Brookhart, 2009).

Teaching and Assessing Developmental Learning Targets One way to begin designing instruction and assessing progress toward developmental objectives is to list several specific learning targets for each objective. The targets should represent the *key* performances expected of a student at a particular grade level. This is illustrated in the following example, which clarifies a broad instructional objective in science by listing several specific learning targets that support it:

Example

(based on Klopfer, 1969)

Developmental learning target:	Interprets and uses Boyle's law to explain phenomena and solve problems.

Specific learning targets clarifying this developmental target:

1. States a definition of Boyle's law.
2. States the domain to which Boyle's law applies.
3. Describes the relationship between Boyle's law and Charles's law.
4. Uses Boyle's law to explain an observation in a lab experiment.
5. Appropriately analyzes a new (to the student) situation in terms of Boyle's law.
6. Solves a new problem or makes an appropriate choice for a course of action, taking into account the implications of Boyle's law.

Although this list of six specific objectives might be made longer, the six objectives would likely be considered adequate for describing what is meant by "interpreting and using Boyle's law" at the end of an introductory course in high school physics. Specific tasks could then be prepared for assessing achievement of the six specific objectives. Some tasks might assess only one of these learning targets; others could require a student to use several of these learning targets in combination. A student's overall score could be interpreted as indicating the degree to which a student has acquired the ability to interpret and use Boyle's law, rather than as a "mastery/nonmastery" description.

SOURCES FOR LOCATING LEARNING TARGETS

You may find lists of learning targets in instructional materials and teachers' manuals, local and state curriculum frameworks, state Websites containing performance standards, reports of the National Assessment of Educational Progress, books on teaching methods, manuals accompanying tests (especially criterion-referenced tests), and reports from educational associations. More than likely you will have to adapt the learning targets you find in these sources to your own situation. Nevertheless, these sources do provide a starting place: It is much easier to adapt and revise learning target statements than to write them without any assistance.

Also, a learning target often will cut across several lessons or subject areas. The ability to use library and print resources to obtain information for a report, for example, is likely to be a learning target common to social studies, mathematics, and language arts curricula. The taxonomies in the next section were created so that each category would apply across several curricular areas.

TAXONOMIES OF LEARNING TARGETS

Simply writing learning targets "off the top of your head" can be frustrating because a seemingly endless number of possible targets exist. Further, if you are unaccustomed to writing learning targets, you are likely to write first those targets that have a very narrow focus, specify content topics, and represent lower level cognitive skills. A taxonomy can help you bring to mind the wide range of important learning targets and thinking skills.

Taxonomies of instructional learning targets are highly organized schemes for classifying learning targets into various levels of complexity. Generally, educational learning targets fall into one of three domains, although a complex performance may involve more than one of them.

1. **Cognitive domain:** Targets focus on knowledge and abilities requiring memory, thinking, and reasoning processes.
2. **Affective domain:** Targets focus on feelings, interests, attitudes, dispositions, and emotional states.
3. **Psychomotor domain:** Targets focus on motor skills and perceptual processes.

Learning targets within each domain may be classified by using a taxonomy for that domain. Because there is more than one way to define a classification scheme, several different taxonomies have been developed for sorting learning targets in a given domain. Only two of these taxonomies for the cognitive domain are described here. Other cognitive domain taxonomies are summarized in Appendix D. Chapter 6 will discuss using taxonomies to develop an assessment plan. The other chapters in Part II discuss creating tasks to assess learning targets at different taxonomy levels.

COGNITIVE DOMAIN TAXONOMIES

Bloom's Taxonomy

The *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain* (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956) had an enormous influence on how we think of educational goals and on teaching practice. We summarize it here because it is still used. This taxonomy is a comprehensive outline of a range of cognitive abilities that you might teach, classified into six major headings arranged from simple to complex.

The six main headings of the original Bloom's taxonomy are Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation.

1. **Knowledge** involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. For measurement purposes, the recall situation involves little more than bringing to mind the appropriate material. (p. 201)
2. **Comprehension** represents the lowest level of understanding. It refers to a type of understanding or apprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating it to other material or seeing its fullest implications. (p. 204)
3. **Application** involves the use of abstractions in particular and concrete situations [to solve new or novel problems]. The abstractions may be in the form of general ideas, rules of procedure, or generalized methods. The abstractions may also be technical principles, ideas, and theories, which must be remembered and applied. (p. 205)
4. **Analysis** involves the breakdown of a communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear and/or the relations between the ideas expressed are made explicit. Such analyses are intended to clarify the communication, to indicate how the communication is organized, and the way in which it manages to convey its effects, as well as its basis and arrangements. (p. 205)
5. **Synthesis** involves the putting together of elements and parts so as to form a whole. This

involves the process of working with pieces, parts, elements, and so on, and arranging and combining them in such a way as to constitute a pattern or structure not clearly there before. (p. 206)

6. **Evaluation** requires judgments about the value of material and methods for given purposes, quantitative and qualitative judgments about the extent to which materials and methods satisfy criteria, and the use of a standard of appraisal. The criteria may be those determined by the student or given to him. (p. 207)

Revised Bloom's Taxonomy

Relationship of the Revision to the Original The original *Taxonomy of Educational Objectives* is still in wide use in schools, but has been revised as *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Anderson et al., 2001). The revised taxonomy improves on the original by adding a two-dimensional framework into which you may classify learning targets and assessment items. The two dimensions are the Knowledge Dimension and Cognitive Process Dimension.

The Cognitive Process Dimension is very much like the original Bloom's *taxonomy*. Its categories are Remember, Understand, Apply, Analyze, Evaluate, and Create. The cognitive processes of Synthesis and Evaluation from the old taxonomy have switched their order and become Evaluate and Create in the new taxonomy. This makes sense, in that evaluation requires making a judgment after analyzing something against criteria, and creating requires putting together something new. The definitions of the Cognitive Process Dimension categories remain like the original Bloom's taxonomy definitions presented above, with one exception: Knowledge.

Bloom's original Knowledge category has been divided into two parts: the Knowledge Dimension and the Cognitive Process category Remember. The Knowledge Dimension has four subcategories: Factual Knowledge, Conceptual Knowledge, Procedural Knowledge, and Metacognitive Knowledge. The Knowledge Dimension contains the type of content a learning target refers to: a fact, a concept, a procedure, or a metacognition. More information about metacognition is presented in Appendix F.

1. **Factual Knowledge**—This category of learning targets asks students to learn facts.
2. **Conceptual Knowledge**—This category of learning targets asks students to learn ideas, generalizations, and/or theories.
3. **Procedural Knowledge**—This category of learning targets asks students to demonstrate procedures or ways of doing things.
4. **Metacognitive Knowledge**—This category of learning targets asks students to be aware of and understand what they know.

The Taxonomy Table A two-dimensional table, the Taxonomy Table is constructed to describe the location of a learning target and its corresponding assessment on both dimensions simultaneously (see Figure 2.3). The figure shows 24 cells, each defined by one knowledge and one cognitive process subcategory. Note that the subcategories of the Knowledge Dimension are lettered, whereas the subcategories of the Cognitive Process Dimension are numbered. As a shortcut, we can refer to a particular cell by its letter and number. Thus, a learning target that requires students to remember some factual knowledge is placed in cell 1A.

Classifying Learning Targets and Assessment Items Suppose you are teaching students to understand the elements that authors use when writing short stories. Suppose the short stories you select all concern people's personal problems, and that the characters in these stories handle their personal problems inappropriately. The learning targets and questions that follow may be used to help you direct your assessment plans. Later chapters will detail how to design assessment tasks. At this point we are studying only the range of thinking skills that should be taught and assessed. Also remember that the examples are classified into the most appropriate cell(s) in the taxonomy, and that they may also overlap into some of the other cells.

Example

Remember Factual Knowledge [1A]

Sample learning target:	Recall the main characters in each of the short stories read and what they did.
Sample assessment items:	(1) List the names of all of the characters in the *Witch's Forest*. (2) In the *Witch's Forest*, what did Sally do when her mother refused to let her go into the forest?

FIGURE 2.3 **Taxonomy Table from the revised taxonomy.**

	Cognitive Process Dimension					
Knowledge Dimension	**1. Remember**	**2. Understand**	**3. Apply**	**4. Analyze**	**5. Evaluate**	**6. Create**
A. Factual Knowledge						
B. Conceptual Knowledge						
C. Procedural Knowledge						
D. Metacognitive Knowledge						

Source: From Lorin W. Anderson & David R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing* © 2001. Published by Allyn and Bacon, Boston, MA. Copyright © 2001 by Pearson Education. Reprinted by permission of the publisher.

Example

Understand Conceptual Knowledge [2B]

Sample learning target: Explain the main ideas and themes of the short stories that we read.

Sample assessment item: Write using your own words what the *Witch's Forest* was all about.

Example

Apply Conceptual Knowledge [3B]

Sample learning target: Relate the personal problems of the characters in the short stories that we read to problems that real people face.

Sample assessment item: Are the problems Sally had with her mother in the story similar to the problems you or someone you know has with his or her mother? Explain why or why not.

Example

Analyze Procedural Knowledge [4C]

Sample learning target: Identify the literary devices that authors use to convey their characters' feelings to the reader.

Sample assessment item: In *Witch's Forest,* Sally was upset with her mother. In Dog Long Gone, Billy was upset with his brother. What words and phrases did the authors of these two stories use to show how upset these characters were? Explain and give examples.

Example

Evaluate and Create using Conceptual and Procedural Knowledge [5B,C;6B,C]

Sample learning targets:

(1) Develop one's own set of three or four criteria for judging the quality of a short story. [6B,C]

(2) Use the three or four criteria to evaluate several new stories that were not read in class. [5B,C]

Sample assessment items:

(1) So far we have read four short stories. What are three or four different traits that make a story high quality? Use these traits to develop three or four criteria that you could use to evaluate the quality of any short story.

(2) Read the two new short stories assigned to you. Use the criteria you developed to evaluate these two stories. Evaluate each story on every criterion. Summarize your findings.

Example

Create using Conceptual and Procedural Knowledge [6B,C]

Sample learning target: Describe, across all of the stories read, the general approach that the characters used to resolve their problems unsuccessfully.

Sample assessment item: So far we have read *Witch's Forest, Dog Long Gone, Simon's Top, and Woman With No Manners*. In every story one character was not able to solve the personal problem he or she faced. What were the ways these characters tried to solve their problems? What do these unsuccessful ways to solve problems have in common?

Figure 2.4 shows how learning targets in science and social studies may be classified in the revised taxonomy. The value of such a taxonomy is that it calls your attention to the variety of abilities and skills toward which you can direct instruction and assessment.

Advantage of the Revised Taxonomy The advantage of the revised taxonomy is that it allows you to consider a broader range of learning targets than the original one-dimensional taxonomy. If you classify your learning targets, your assessment items, and your teaching activities into the Taxonomy Table shown in Figure 2.3, you can immediately see the types of knowledge and thinking on which your instructional unit focuses. Not every unit should have learning targets and assessments in every one of the 24 cells, of course. But over the semester, your teaching should address and evaluate students' learning in all (or nearly all) of them.

Different Modes of Assessments for Different Taxonomy Levels

Note that learning targets classified in the first three cognitive categories are more easily assessed with short-answer, true-false, multiple-choice, or matching test items. Learning targets classified in the last three cognitive categories might be partially tested by such item formats, but their assessment usually requires a variety of other procedures such as essay questions, class projects, observing performance in labs, and portfolios. Learning targets at more

FIGURE 2.4 **How different outcomes for science and social studies may be classified using the Anderson et al. revised taxonomy.**

Anderson et al. Category	Science		Social Studies	
Remember	• Recall the names of parts of a flower • Identify and label the parts of insects • List the steps in a process	1A 1A 1C	• List known causes of the Civil War • Recall general principles of migration of peoples of Africa	1B 1B
Understand	• Explain the digestive processes in one's own words	2B	• Explain the meaning of technical concepts in one's own words • Give examples of propaganda usage from current events	2B 2B
Apply	• Use scientific principles to make a simple machine • Find real examples of igneous rock and mineral formations	3B,C 3B	• Use specified principles to explain current events • Carry out a survey and collect data from the field	3B,C 3C
Analyze	• Show how scientific principles or concepts are applied in the design of a refrigerator	4B,C	• Identify the credible and noncredible claims of an advertisement for clothing • Show the different component parts of a political speech	4B,C 4B
Evaluate	• Use criteria or standards to evaluate the conclusions drawn from the research findings	5B,C	• Use a specific set of criteria to evaluate several political speeches	5B
Create	• Determine what the rule is that underlies the results obtained from several experiments or investigations	6B	• Show the similarities among several schools of social thought • Develop plans for peace between two countries	6B 6B,C

complex thinking levels require students to actually produce or create something, rather than simply to answer questions. Carefully reading the various subcategories of the taxonomy in Appendix D should make this more apparent.

Condensing the Taxonomy Because it is sometimes difficult for teachers to classify their learning targets into all six cognitive categories, some schools have opted to use a shorter version of the Bloom taxonomy. For example, some have reduced it to three categories: Remember (or Knowledge), Understand (or Comprehension), and Higher-Order Thinking. The "Higher-Order Thinking" category collapses Apply, Analyze, Evaluate, and Create learning targets into one group. Other schools formed three categories somewhat differently: Remember and Understand (Knowledge and Comprehension), Apply, and Higher-Order Thinking (including Analyze, Evaluate, and Create). The advantage of these condensations is that they eliminate the need for struggling with how to classify learning targets into one of the top three categories of the taxonomy.

A disadvantage of using a condensed version of the taxonomy is that teachers may stop trying to teach learning targets in the Evaluate and Create categories. Because, after condensing, Apply and Analyze will be in the same category as Evaluate and Create, it is easy to avoid making the necessary distinctions among these four. As a result, a teacher may settle for not including the highest two categories of learning targets in lesson and assessment plans.

Problems When Classifying Learning Targets Using a Taxonomy

Taxonomies are not teaching hierarchies. Their only purpose is to classify various learning targets and assessment tasks. For example, you should not teach "knowledge" first and "comprehension" second. If you did that, younger and lower-achieving students would be doomed to spend all their time on drill. Use the taxonomy to help you explore each learning goal at several levels.

It is also important to recognize that student performance on complex tasks involves using several thinking skills at the same time. It is possible, therefore, to classify a given learning target or assessment task into more than one taxonomy category. The authors of the revised taxonomy, in fact, encourage you to do so (Krathwohl, 2002).

For most classroom purposes, classify each learning target into the category that represents the thinking skill that is (a) most prominently used by or (b) the main intent of the learning target or assessment task. Then use the classification to decide if some important skills have received too little or too much attention in your teaching and assessment.

The main purpose for suggesting that you use a taxonomy for assessment is to give you a tool to judge whether you have taught and assessed a wide enough range of higher- and lower-order thinking skills. A taxonomy may help you find the gaps. Including a wide range of thinking skills in an assessment usually improves its validity.

Choosing a Taxonomy

We have discussed two different schemes for classifying cognitive learning targets. There are many more taxonomies or schemes that we have not discussed, some of which are in Appendix D of this book. Which one should you use? That depends on whether this is a personal decision for use in your classroom only or a more general decision about a taxonomy that will be used throughout your school system.

✔ **Checklist**

Criteria for Selecting a Taxonomy of Cognitive Learning Targets

1. *Completeness:* To what degree can your major learning targets be classified within this taxonomy?

Not at all   Somewhat   To a great extent

2. *Point of view:* To what extent can this taxonomy be used as a platform for explaining your teaching methods or your curriculum characteristics to others?

Not at all   Somewhat   To a great extent

3. *Reform:* To what extent can this taxonomy help you evaluate your curriculum or your learning targets and lead you to revise the learning targets?

Not at all   Somewhat   To a great extent

4. *Simplicity:* How easy is it for parents, teachers, and education officials to understand this taxonomy?

Not at all   Somewhat   To a great extent

5. *Reporting:* How useful is this taxonomy in organizing reports on assessment results for individual students, educational officials, government officials, or the public?

Not at all   Somewhat   To a great extent

To choose among the various taxonomies, apply the practical criteria in the checklist to judge each taxonomy or classification scheme you are considering. If the decision is a personal one for a single classroom, then not all criteria may apply.

EVALUATING THE LEARNING TARGETS OF A COURSE OR UNIT

Never teach a unit until you list the learning targets for that unit. It is important to develop a complete or comprehensive list of learning targets. A complete list is not necessarily long, however. You can use this checklist to evaluate your list of learning targets:

✔ **Checklist**

Checklist for Evaluating a List of Learning Targets for a Course or Unit

1. Are all the learning targets appropriate for students' educational level?
2. Is the list of learning targets limited to only the important outcomes for the course or unit?
3. Are all the learning targets consistent with your state's published learning standards?
4. Are all the learning targets consistent with your local school's philosophy and general goals?
5. Can all the learning targets be defended by currently accepted learning principles?
6. Can all the learning targets be taught within the time limits of the course or unit?
7. Can all the learning targets be taught with available teaching resources?

HOW TO WRITE SPECIFIC LEARNING TARGETS

To be useful for classroom instruction and assessment (Gronlund & Brookhart, 2009) learning targets must be:

1. **Student centered:** Learning targets should focus on the student.
2. **Performance centered:** Learning targets should be worded in terms of what a student can perform after the required learning experiences.
3. **Content centered:** Learning targets should state the specific content to which the student should apply the performance.

Student Centered

Because instruction focuses on changes in student performance, learning targets should describe student performances. It is not unusual, however, for some teacher guides, curriculum frameworks, and other materials to contain statements that do not focus on the student. Consider this statement:

Example

Poor: Provide the opportunity for students to express their opinions in classroom discussions about why peace is so difficult to attain.

The problem with the preceding statement is that it is an activity statement *for teachers* rather than a learning target for students. You may "provide the opportunity for students to express their opinions," yet each student may not express his or her opinion. Learning targets need to be student centered if they are to be the basis for crafting assessment procedures. Thus, you should say:

Example

Better: A student will express his or her opinion in classroom discussions about why peace is so difficult to attain.

Student-centered learning targets allow you to decide whether the students actually have achieved what you intended from the lesson.

Performance Centered

Not only should a learning target refer to a student, it should state a performance—that is, an observable activity. This can be accomplished by being sure that the statement includes an action verb that specifies a student performance.

To help beginners write learning targets that describe students' performances, Figure 2.5 lists further examples of various action verbs. These verbs are organized according to the cognitive dimension of the Anderson et al. revised taxonomy. When verbs such as these are used in statements of learning targets, the learning targets will usually satisfy the second criterion of expressing observable student performance.

A balance is necessary between verbs that are too broad (and thus imply too many nonequivalent performances) and those that are too specific

FIGURE 2.5 **Action verbs to use when writing learning targets.**

Remember	Define, describe, explain, identify, label, list, match, name, outline, reproduce, select, state
Understand	Convert, describe, distinguish, estimate, extend, generalize, give examples, paraphrase, rewrite, summarize
Apply	Apply, change, classify (examples of a concept), compute, demonstrate, discover, modify, operate, predict, prepare, relate, show, solve, use
Analyze	Analyze, arrange, associate, compare, contrast, infer, organize, solve, support (a thesis)
Evaluate	Appraise, compare, conclude, contrast, criticize, evaluate, judge, justify, support (a judgment), verify
Create	Classify (infer the classification system), construct, create, extend, formulate, generate, synthesize

(and which are often just ways of marking answers). Consider this learning target, which is stated too specifically:

Example

Poor: The student is able to put an X on the picture of the correct geometric shape (circle, triangle, rectangle, square, or ellipse) when the name of the shape is given.

The main intent of such an objective is to select or identify the correct shape, not just to make Xs. Any response that indicates the student has correctly identified the required shape is acceptable. Thus, the learning target should be written as:

Example

Better: The student is able to identify a picture of a geometric shape (circle, triangle, rectangle, square, or ellipse) when the name of the shape is given.

Figure 2.6 suggests some verbs that maintain this balance and illustrates other verbs that are too specific or too broad to make useful learning target statements.

FIGURE 2.6 **Action verbs sometimes used in learning targets.**

Specific but acceptable verbs

add, total
alphabetize
choose
complete, supply
construct, make
convert
count
delete
describe
divide
draw
explain
identify
label
list
match
measure
multiply
name
order, arrange
pick out
regroup
rename
rephrase
select
sort, classify
state
subtract, take away
weigh

Too broad, unacceptable verbs

apply
deduce
do
examine
generate
infer
interpret
observe
perform
respond
test
use

Too specific, essentially indicator verbs

check
circle
color the same as
draw a line between
draw a ring around
put a box around
put a mark on
put an X on
shade
underline
write the letter of
write the number of

Toss-up verbs, requiring further clarification

answer
collect, synthesize
compare
contrast
demonstrate
determine
differentiate
discriminate
distinguish
give
locate
predict

Source: From "Criteria for Stating IPI Objectives" by C. M. Lindvall, 1976, pp. 214–215. In D. T. Gow (Ed.). *Design and Development of Curricular Materials: Instructional Design Articles* (Volume 2). Pittsburgh, PA: University of Pittsburgh, University Center for International Studies.

Content Oriented

A learning target should also indicate the content to which a student's performance is to apply. The following learning target is poor because it lacks a reference to content:

Example

Poor: The student is able to write definitions of the important terms used in the text.

To modify this learning target you need to include a reference to a specific list of "important words" or in some other way describe them:

Example

Better: The student is able to write definitions of the terms listed in the "Important Terms" sections of Chapters 1–5 of the textbook.

If you do not refer to content in your learning target statement, you will be uncertain whether an assessment task is valid for evaluating the student. For example, the assessment may require students to define words that, although in the text, may be unimportant. Without knowing the content, it is difficult for anyone to determine what, if anything, was learned.

ALIGNING ASSESSMENT TASKS WITH LEARNING TARGETS

Chapters 6 through 14 discuss the details of creating high-quality assessments. Here we wish simply to point out that the basic purpose of any assessment is to determine the extent to which each student has achieved the stated learning targets. Although this purpose sounds straightforward, it is not always an easy criterion to meet. The validity of your assessment results determines the quality of your evaluation. Validity has many aspects (see Chapter 3); here we discuss validity only in relation to matching or aligning assessment tasks to learning targets.

Aligning Assessments to Mastery Learning Targets

The specific tasks or procedures you use in an assessment should require the student to display the skill or knowledge stated in the learning target. For instance, if the main intent of your learning target is for a student to build an apparatus, write a poem, or perform a physical skill, your assessment procedure must give the student the opportunity to *perform*. Assessment procedures that require a student only to name the parts of an apparatus, to analyze an existing poem, or to describe the sequence of steps needed for performing a physical skill do not require the performance stated in these learning targets. Therefore, they would be invalid for assessing them: They are not aligned to the learning targets' main intents. A very basic requirement for the validity of classroom assessment procedures is that the assessment procedures should be aligned with the intentions of the specific learning targets that you include in your assessment plan. Methods for developing assessment plans are treated in Chapter 6.

Aligning Assessments to Developmental Learning Targets

As is often the case, developmental learning targets imply a broad domain of performance application. To ensure the validity of your classroom assessment, you may need to assess the same learning target in several different ways. For example, you might assess writing achievement both by scoring several samples of students' written assignments and by using a grammar and usage test. The test provides the opportunity to assess grammar and usage that might not appear in the natural course of the student's writing, but that may well be part of the learning target. Observing a student's natural writing habits permits you to infer how well the student is likely to use language in typical writing situations. Using both procedures increases how comprehensively you assess the student's writing ability and the validity of your evaluation.

Another reason for using more than one assessment procedure is to obtain more reliable results. Your subjective evaluation of a student's written essay on a topic might be supplemented by a test made up of more objectively scored items. Combining the less reliable information about the student's written work (that is, your subjective evaluation) with the more reliable information (the objectively scored test) yields a more reliable overall evaluation result. Reliability is discussed in more detail in Chapter 4.

Aligning Assessments to State Standards

Earlier in this chapter we showed how you can derive your learning targets from your state's standards. It is important that you maintain consistency by aligning your classroom assessments as well as the learning targets with the state's standards. Aligning your assessments with the learning targets that you derive from the state's standards (in the manner we showed earlier) is one way to ensure this. You will want to ensure that your assessments match the span of content covered by the standards, the depth of thinking implied by the standards, the topical emphasis in the standards, and the same types of performances as are specified in the standards.

CONCLUSION

Well-conceived learning targets are the foundation for both instruction and assessment. They are also the means by which instruction and assessment are coordinated. Such coordination or alignment is the basis for valid classroom assessment. In Chapter 3, we consider the broader concept of validity for both classroom and large-scale assessment.

EXERCISES

1. Write three specific learning targets for a lesson you plan to teach. Explain how each learning target meets the three criteria: student centered, performance centered, and content centered.
2. Following are three learning targets. Decide whether each is a mastery learning target or a developmental learning target. Explain your choices.
   a. The student is able to take the square root of any number using a handheld calculator.
   b. The student is able to determine whether the thesis of the argument is supported adequately.
   c. When given data, the student is able to construct a graph to describe the trend in the data.
3. a. Obtain a copy of your state's (or neighboring state's) standards. Analyze the suitability of these statements (i) for planning units and lessons and (ii) for developing assessment exercises. Prepare a criticism of these standards from your point of view as a teacher of a specific grade and subject. In your criticism, be sure to emphasize assessment-related issues. Hint: Log on to your state education department's Website and search for links to the education standards.
   b. Select one unit you are teaching or plan to teach in the future that is based on one or more of your state's standards. Explain what you would need to do to align your classroom learning targets and your student assessments with the state's standards. Summarize the results and report them to your class.
4. Decide whether each learning target listed here belongs to the cognitive, affective, or psychomotor domain. Does the performance of each learning target require some use of elements from domains other than the one into which you classified it? Which one(s)? Explain why. Does this mean you should reclassify that learning target? Explain.
   a. The student is able to tune a television set to get the best color resolution.
   b. The student demonstrates knowledge of parliamentary law by conducting a meeting without violating parliamentary procedures.
   c. The student contributes to group maintenance when working with classmates on a science project.
   d. The student makes five baskets in 10 attempts on the basketball court while standing at the foul line.

Planning for Integrating Assessment and Instruction

KEY CONCEPTS

1. Good assessment planning and good instructional planning are two sides of the same coin; do them together.
2. Formative and summative assessment both require planning.
3. Assessment planning for a marking period should be based on learning goals and outline the main instructional and assessment strategies you will use.
4. Assessment planning for a unit of instruction should be based on learning goals and objectives and detail the instructional and assessment strategies you will use.
5. Use a pretest to help plan your teaching.
6. Differentiated instruction relies on accurate, timely assessment in order to be effective.
7. Use a blueprint to plan individual summative assessments.
8. Use criteria to improve the validity of assessment plans.
9. A wide range of assessment options are available: paper-and-pencil, performance, long-term assignment, and personal communication formats.
10. Assessment for Response to Intervention (RTI) involves planning for screening and for progress monitoring.

IMPORTANT TERMS

assessment planning
best works portfolio
blueprint or table of specifications
criteria for evaluating a planned assessment
differentiated instruction
elements of a complete test plan
equivalence
feedback to students
formative uses of assessment
growth portfolio
informal assessment techniques
marking period
paper-and-pencil assessments
performance assessment
preinstruction unit assessment framework
process
product
progress monitoring
Response to Intervention (RTI)
sizing up
summative uses of assessment
task formats
teaching actions after assessing
unit of instruction

From Chapter 6 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

Plans for teaching are incomplete unless they contain plans for assessment. This chapter focuses on how to improve your **assessment planning**. Good assessment planning and good instructional planning are two sides of the same coin; do them together. Both begin with, and are based on, identifying your learning goals, objectives, or targets—the essential knowledge and skills you want students to learn.

HOW MAKING YOUR OWN ASSESSMENTS IMPROVES YOUR TEACHING

You can expect the following benefits to your teaching as your assessment skills improve.

1. *Knowing how to choose or to craft quality assessments increases the quality of your teaching decisions.* Assessing how your students use their knowledge and skills allows you to monitor and evaluate their progress and appropriately differentiate instruction.
2. *What and how you assess communicates in a powerful way what you really value in your students' learning.* For example, you may tell your students how important it is for them to be independent and critical thinkers, but if your assessments consist of only matching exercises based on facts from the textbook, students will know differently. On the other hand, if your assessments require students to integrate their knowledge and skills to solve "real-life" problems, they learn that you really do expect them to develop integrating and problem-solving abilities.
3. *When you carefully define assessment tasks, you are clarifying what you want students to learn.* To teach effectively you must clearly have in mind how students should demonstrate their achievement. Creating assessment tasks requires you to create situations in which students can demonstrate their achievement.
4. *Learning to create your own assessment tasks increases your freedom to design lessons*. Knowing how to assess students validly, especially in relation to higher-order thinking skills, means that you are no longer chained to the assessment procedures already prepared by textbook publishers and others. Therefore, you can use a wider variety of teaching strategies.
5. *You will improve the validity of your interpretations and uses of assessment results.* Research shows that teachers who have studied assessment, either through coursework or in-service training, are able to recognize and produce better assessments (Boothroyd, McMorris, & Pruzek, 1992; Plake, Impara, & Fager, 1993).
6. *You will improve your appreciation of the strengths and limitations of each type of assessment procedure.* You will be able to use multiple, complementary measures to get a clear picture of what your students know and can do.

ARE YOU ASSESSING FOR FORMATIVE OR SUMMATIVE PURPOSES?

Formative and summative assessment both require planning. Both should be based on the same learning targets. Figure 6.1 shows common uses for classroom assessment results. The uses are organized into two groups: formative and summative. One use of assessment, controlling students' behavior, is not listed in Figure 6.1 because it is a poor, and sometimes unethical, practice. Controlling students through assessments turns a process of information gathering into a process of threatening and punishing, with negative consequences for learning and self-efficacy.

Formative Purposes of Assessment

Formative uses of assessment help you guide or monitor student learning while it is still in progress. High-quality formative assessment and **feedback to students** increases student learning (Black & Wiliam, 1998). In general, formative assessments are less formal. We recommend that you record the results of these assessments to help your memory; however, you do not use them to report official letter grades or achievement progress. (As you can see from Figure 6.1, each of the formative uses helps you plan what and how to teach.)

Typically, you use the most informal assessments for sizing-up purposes. **Sizing up** means to form a general impression of a student's strengths, weaknesses, learning characteristics, and personality at the beginning of a course or at the start of the year. The following example illustrates how a teacher pulled together various informally obtained pieces of information to size up Saleene, a fifth-grade student:

> Saleene (a fifth grader) walks into class each day with a worried and tired look on her face.

FIGURE 6.1 **Examples of basic purposes for which classroom assessment results are used.**

I. *Formative uses* help teachers monitor or guide student learning while it is still in progress.
   - A. Sizing-up uses help a teacher form initial impressions of students' strengths, weaknesses, learning characteristics, and personalities at the beginning of the year or course.
   - B. *Diagnosing individual students' learning needs* helps a teacher and the student identify what the student has learned and what still needs to be learned, decide how instruction needs to be differentiated, and decide what feedback each student needs about how to improve.
   - C. *Diagnosing the group's learning needs* helps a teacher identify how the class as a whole has progressed in its learning, what might need to be reinforced or retaught, and when the group is ready to move on to new learning.
   - D. *Using assessment procedures as teaching tools* is a way in which a teacher uses the assessment process as a teaching strategy. For example, a teacher may give practice tests or "mock exams" to help students understand the types of tasks used on the assessment, practice answering and recording answers in the desired way, or improve the speed at which they respond. In some cases, the performance assessed is identical or nearly identical to the desired learning target so that "practicing the assessment" is akin to teaching the desired learning target.
   - E. *Communicating achievement expectations* to students is a use in which a teacher helps clarify for students exactly what they are expected to be able to perform when their learning is complete. This may be done by showing the actual assessment tasks or by reviewing the various levels or degrees of performance of previous students on specific assessment tasks so that current students may be clear about the level of learning expected of them.
   - F. *Providing specific feedback* gives students information about how to improve.
   - G. *Promoting students' self-assessment* helps students monitor their own learning, set goals, and take action to meet them.
   - H. *Planning instructional uses* helps a teacher design and implement appropriate learning and instruction activities, decide what content to include or emphasize, and organize and manage the classroom as a learning environment.

II. *Summative uses* help a teacher evaluate student learning after teaching one or more units of a course of study.
   - A. *Assigning grades for report cards* is a way in which a teacher records evaluations of each student's learning progress to communicate evaluations to students, their parents, and responsible educational authorities.
   - B. *Placing students into remedial and advanced courses* is a way in which a teacher attempts to adapt instruction to individuals' needs when teaching is group based. Students who do poorly in the teacher's class may be placed into remedial classes that provide either alternate or supplemental instruction that is more suitable for the students' current level of educational development. Similarly, students whose educational development in the subject is above that of the rest of the class may be placed into a higher level or more enriched class.
   - C. *Evaluating one's own teaching* requires a teacher to review the learning that students have been able to demonstrate after the lessons are complete, identify which lessons were successful with which students, and formulate modifications in teaching strategies that will lead to improved student performance the next time the lessons are taught.

> Praising her work, or even the smallest positive action, will crack a smile on her cheeks, though the impact is brief. She is inattentive, even during the exercises we do step by step. Saleene has a hearing disability that makes it hard for her to follow directions and classroom discussions. She is shy, but sometimes will ask for help. But before she gives herself a chance, she will put her head down on her desk and close her eyes. Her self-esteem is low. I am concerned that she will be this way all year. (Airasian, 2001, p. 38)

You can see that this teacher used information about Saleene's cognitive, affective, and psychomotor traits to help form a general strategy for how to teach her.

Other formative decisions also require quality information. These include diagnosing individual students' learning needs, communicating achievement expectations, using assessment in instruction, diagnosing the group's learning needs, providing feedback, promoting student self-assessment, and planning instruction (see Chapter 7). These decisions require valid information from carefully planned assessment.

Summative Purposes of Assessment

Summative uses of assessment help you evaluate your students and your own teaching after you finish teaching one or more units. Often we use summative information about students' achievement to count toward their grades for a marking period (see Chapter 14). Parents and school authorities interpret those grades as the progress students have made toward achieving the curriculum's learning targets. Figure 6.1 also notes that placement and evaluation decisions are summative uses for assessment results. Because of the finality of summative assessment, you should prepare to keep records of students' results on summative assessments and ensure the validity of each result for supporting the decisions based on them.

ASSESSMENT PLANNING FOR A MARKING PERIOD

Keeping in mind that you need to plan for both formative and summative assessment, the next thing to consider is the period for the plan. You may plan for a year, a semester, a marking period, a unit, or a lesson. Your plans for larger/longer segments of your teaching will be less detailed than your plans for smaller/shorter segments.

Plans for a year or a semester set out the general approaches and strategies you will use to teach and to assess. Such a plan contains an outline of the topics you will teach, the general learning targets your students will achieve, and the main strategies you will use to assess them.

Plans for a marking period usually apply to two or three units of instruction. A **marking period** is the number of weeks you must teach before you need to prepare a grade for each student's report card. In a typical academic year a marking period consists of 9 weeks. A **unit of instruction** is a teaching sequence covering from 1 to 7 weeks of lessons, depending on the students and topics you are teaching. You use plans for instructional units to break down and organize the larger curriculum into manageable teaching, learning, and assessment sequences. Planning for several units at one time allows for sequencing the units and for keeping your teaching and assessment approaches consistent. It also allows you to describe your plans for formative and summative assessment.

Plans for only one unit will necessarily be more detailed. You will describe the specific content, concepts, procedures, terminology, and thinking skills your students will learn and use. You also describe your teaching activities and your students' learning activities. You identify the learning targets of the lessons, the specific formative and summative assessments you will use, and when you will use them.

The shortest term for planning is for a single day or one lesson. As you teach, you will begin to reflect on what you have previously taught these students and how well your students have achieved the unit's learning targets to date. This reflection is an opportunity for you to adjust your unit plan. Your teaching and assessment strategies become more fine-tuned, adapting to your students' achievement. Each day, you adjust your teaching as you gather new information about your students and your teaching.

This latter point illustrates that your teaching and assessment plans are not set in stone. They are guidelines for teaching and assessing. They are flexible and subject to change as new information about your students' achievements accumulates.

Example of How to Develop an Assessment Plan for a Marking Period

Assessment planning for a marking period should be based on learning goals and outline the main instructional and assessment strategies you will use. Because this is an assessment book we shall emphasize the assessment aspects of planning, but your planning also will include instructional ties. Suppose you are teaching middle school science. Suppose, further, that you are planning for a 9-week marking period. Perhaps you plan to teach two units: one on the water cycle and one on weather and weather systems. For each unit you would outline the major points of content you will cover, the general sequence and timing of the units, and, most important, the learning targets your students will achieve from each unit.

On the teaching side, you will need to answer a variety of questions. What overall approach and teaching strategy will you adopt? The water cycle and weather units are related; how will you make that clear to students? What kinds of learning activities will you need to create and use (e.g., creating a demonstration of condensation, cloud simulation, building a diorama of the water cycle, drawing weather maps, measuring variables related to weather such as wind speed and precipitation, collecting and reading weather maps, or conducting a weather prediction activity)?

Part of your teaching plan must include student evaluation. How will you evaluate students' achievement of the learning targets? What are your general strategies for formative evaluation? Perhaps you plan for some in-class activities and exercises that will allow you to evaluate how well students are progressing. These also allow you to give students appropriate feedback. Perhaps you plan homework exercises. These allow you to evaluate whether students have mastered the basic concepts. Your thinking should include planning for how often you assess. At what points in the lessons will homework or quizzes be appropriate, for example?

To provide formative feedback to students, you will have to evaluate their work. Will students and/or their peers evaluate performance? If so, students will need evaluation criteria and scoring rubrics. When you use oral questioning, what levels of the taxonomy will you emphasize most? How will you respond to intermediate steps toward larger projects (plans, outlines, drafts, etc.)?

In order for formative feedback to help students improve their learning, you will probably have to teach your students how to use this feedback and provide opportunities for them to do so. You may need to teach them how to review and evaluate their own work as they proceed through the lessons.

Your summative evaluation strategy also needs to be planned. You might use a paper-and-pencil test at the end of each unit. You might use a project for one unit and a performance activity for another. For example, students may collect weather data and use those data to predict the weather. For some other subjects, term papers, independent investigations, or portfolios might prove useful for summative evaluation. You will want to build in formative assessment opportunities along the way for the larger projects.

Your plan must include the weighting of each component as part of a final grade: How much will the tests, homework, projects, and so forth count toward the grade? Will each count equally, or will some weigh more heavily than others? To be fair, you will need to explain the weighting to students in advance.

Example of an Assessment Plan for a Marking Period

Figure 6.2 shows an assessment plan that a hypothetical teacher created when teaching the two science units referred to in the preceding paragraphs. Your own plan may be handwritten or word processed and used as a working document as you teach. The main points are that by planning you have (a) decided ahead about when and how you will assess, (b) recorded this thinking so that you do not forget, and (c) followed a systematic plan to achieve your assessment goals.

ASSESSMENT PLANNING FOR ONE UNIT OF INSTRUCTION

Assessment planning for a unit of instruction should be based on learning goals and objectives and detail the instructional and assessment strategies you will use. You should be able to explain why you need to use each assessment strategy, how the assessments are related to the learning targets and the lessons, and what actions you will take once you have information about the students' achievement.

Example of an Assessment Plan for One Unit

Figure 6.3 shows an example of an assessment plan for one of the science units in Figure 6.2. Keep in mind that it includes all the thinking a teacher might use when deciding what assessments to conduct. Your own plan might not be so detailed because the thinking remains in your head. The important points are that you can explain when and why you are using different assessment methods, that you match the assessment methods with the learning target(s) for which they are appropriate, and that you can state what teaching action you will take once the information is gathered. *Assessments are useless if you do not take action when you see the results*.

Observe how Figure 6.3 is organized. Notice that in this example seven lessons are planned. Directly below each lesson is a brief statement of the lesson's main learning target. The various types or methods of assessment (pretest, observation, homework, quizzes, independent investigations, end-of-unit test) are listed in the far-left column. Notice that as you go down the column, the purposes of assessment become more summative and the assessment procedure becomes more formal.

The statements written in the body of this figure describe the purpose, procedure, and action to be taken for each assessment. The **teaching actions after assessing** are steps the teacher will take to improve students' achievement based on the assessment results.

When the statements in Figure 6.3 are spread across the page, that means the assessment's purpose, procedure, and actions apply to all of the lessons. In the figure, observation, oral questioning, and homework are of this character. Statements that appear directly below one or two lessons

FIGURE 6.2 **A long-term plan for a marking period in which two elementary science units will be taught.**

Unit 1. The Water Cycle	
General learning target:	Understanding what the water cycle is, how it works, and how it helps living things. Ability to explain the water cycle and apply it to real life.
Time frame:	It will take 2 weeks to complete.
Formative assessment:	(a) Three homework assignments (taken from Chapter 8) (b) Condensation demonstrations (Group activity; I will ask students to explain what they are doing, how it relates to the water cycle, and how it relates to real life.) (c) Short quiz on the basic concepts at the end of Week 1
Summative assessment:	A written test at the end of the unit (short-answer and an essay)
Weights:	(a) Homework 10% (b) Quiz 10% (c) End-of-unit test 80%
Unit 2. Weather Systems and Predicting Weather	
General learning target:	Understanding basic weather patterns, their movements, and their influence on local climate. Ability to understand weather maps, weather forecasts; ability to collect weather data and use them to make simple predictions.
Time frame:	It will take 7 weeks to complete.
Formative assessment:	(a) Seven homework assignments (taken from Chapter 9 and my own) (b) Seatwork on drawing a simple weather map with symbols (I will circulate among students and ask questions to check their understanding.) (c) Correct use of simple instruments to gather weather-related data (I will have each student demonstrate each instrument's use and give them feedback when necessary.) (d) Collection of weather maps and forecasts (I will discuss with students what the maps and forecasts mean and be sure they understand them.) (e) Four quizzes on the major concepts and a performance activity (Week 1, Week 3, Week 4, and Week 5)
Summative assessment:	(a) Map drawing (I will provide weather information; students will draw corresponding maps independently.This will be Quiz 4.) (b) End-of-unit test (short-answer, matching, map identification, essay question) (c) Independent investigation (Collect weather data for 2 weeks and make daily 2 day weather predictions. I will structure this activity. It will be done toward the end of the unit.)
Weights:	(a) Homework 10% (b) Quizzes 10% (c) Independent investigation 30% (d) Map drawing 20% (e) End-of-unit test 30%
Marking Period Grade	
	Unit 1 marks count 30% Unit 2 marks count 70%

mean that the assessment applies to only those one or two lessons. The quizzes, independent investigation, and end-of-unit test are of this character. Because the seven lessons are spread out in sequence over time, the plan shows that some assessments occur at different times throughout the unit.

PRETESTING TO PLAN YOUR TEACHING

Notice in Figure 6.3 that the teacher gave a pretest about a week before teaching this unit. The pretest results were not used to grade students. Rather, they were used primarily to help the teacher

FIGURE 6.3 **An assessment activity plan for one unit of instruction.**

More Formative in Nature

Assessment techniques	Description of assessment purpose, activity, and follow-up action (use)						
Pretest	About a week before beginning this unit, I will give a very brief pretest to get a sense of students' attitudes, experiences, knowledge, and belief about weather. (See Figure 6.4.) *Action:* I will use this information to help me develop discussions in class, to develop lessons that overcome students' misconceptions and fears about the weather, and to build on what students already know.						
	Lesson 1 Comprehending basic weather concepts	**Lesson 2** Distinguishing weather patterns and systems	**Lesson 3** Identifying local weather conditions and patterns	**Lesson 4** Using basic tools for measuring weather	**Lesson 5** Understanding and making weather maps	**Lesson 6** Collecting and recording local weather data	**Lesson 7** Using data to predict local weather
Observation and oral questioning	In every lesson, I will observe students and ask questions during the lesson to assess how well they are responding to the material, how well they seem to understand the daily activities and assignments, and whether they have any misconceptions about the weather concepts we are studying. *Action:* I'll adjust my teaching if most of the class is having difficulty. If only a few are experiencing difficulty, I'll work with them individually, in small groups, or ask another student to teach the concept.						
Homework	I will assign homework after every lesson. Homework activities will focus on observing and discovering real-world examples of the weather concepts we learn in class. Students will record their observations and write explanations of them using proper scientific language learned in the unit. *Action:* As I read students' homework responses, I will note for each student how accurately and fluently the student uses scientific language to discuss the weather. I will also evaluate their observational and recording skills.I will reteach those materials for which many students experience difficulty. If only a few are having difficulty, I will work with them individually.						
Quizzes	**Quiz 1** (covers Lesson 1): Short-answer questions testing basic vocabulary *Action:* Students not mastering the basic concepts will be retaught.	**Quiz 2** (covers Lessons 2 and 3):Short-answer questions with some diagrams. Focuses on weather patterns: local, national, and international. *Action:* I will use this quiz to monitor students' understanding of weather patterns and systems. I'll reteach or move on, depending on the outcomes.		**Quiz 3** (covers Lesson 4): This will be a performance activity. I want to be sure each student can use with accuracy the weather-measuring tools and can record data properly. *Action:* I will correct errors on the spot.		**Quiz 4** (covers Lesson 5): I want students to read, interpret, and draw simple weather maps. I will give weather data to the students and ask them to draw an appropriate map using the weather data.I will also give maps already drawn and ask students to interpret them. *Action:* I will reteach if there are problems.	
Independent investigation (performance assessment)				**Predicting the Weather** (begins after Lesson 4, and includes Lessons 5 and 6): This performance assessment will help me evaluate whether students can apply the concepts from the lessons to the real world. It will help me evaluate whether they can synthesize and use criteria to evaluate the data they collect. Students will collect and measure weather data, record it, and use it to predict the local weather for two days in advance. They will repeat the exercise every day for at least 2 weeks. They will work independently. They will prepare a report describing what they did and evaluating their investigation and its accuracy. *Action:* This is a type of summative evaluation. I will use the exercise to help me decide how well the students have learned the concepts and principles in this unit. I should have a pretty good idea whether students can apply what they learned in class.			
End-of-unit test						**Unit Test** (covers all lessons): This will come at the end of all the lessons. It will be a paper-and-pencil test given in class. (I may give it over 2 days.) It will be comprehensive, covering most of the important learning targets in the unit.) *Action:* I will use the results of this test along with the results from homework, quizzes, drawing, and the independent investigation to assign a grade to the students for the unit. (Weights are given in Figure 6.2.)	

More Summative in Nature

understand the students' attitudes, knowledge, beliefs, and experiences about the weather so that the teacher could better teach the unit.

Importance of Preinstructional Unit Assessment

As you plan instruction for a unit, you must consider more than covering the material. In most subjects, students bring to the unit a complex combination of knowledge, experiences, skills, beliefs, and attitudes that are especially related to the topics to be taught. If you understand your students' thinking before teaching them, you can build your instruction on it. The "pretest" does not need to be a formal test. You may, for example, have a class discussion about some of the topics that you will be teaching in an upcoming unit. From this discussion you can gauge how much the class already knows about the topics and what kinds of misconceptions they may have. Use this information to plan your teaching of the unit.

Often students' beliefs about a topic are contrary to what you will teach. Even after you present the information, students' beliefs may not change. If students do not believe what you are teaching, then they do not integrate new concepts into their existing ways of thinking, and they will be unable to apply that information in the future. For example, youngsters know that wearing sweaters keeps them warm. When teaching a science unit on insulating properties, you may teach that air has insulating properties. If you ask youngsters what happens to the temperature of a cold bottle of soft drink when you wrap it in a sweater, many may say it gets very warm. If you tell them it will stay cold, many will not believe you because they know sweaters keep them warm. Knowing this, your teaching will have to include activities that change students' beliefs by building on their prior experiences and knowledge. Your instruction will have to offer a real demonstration and comprehensive explanation—for example, why a sweater keeps the student warm *and* the soft drink cool—before that instruction can alter their beliefs.

A Framework for Constructing Instruments

A **preinstruction unit assessment framework** is a plan you use to help you assess cognitive and affective learning targets of an upcoming unit. Preinstruction assessments should be relatively short, however, so focus your assessment on only a few core elements. Do a written assessment so you can easily summarize the information and use it to make your planning decisions. You could also organize a class discussion around the results.

It is especially helpful if you adopt a set framework and use it to generate assessment questions for every unit you teach. This establishes a comprehensive and consistent approach to gathering and using information. Originally developed for middle school science, the framework in Figure 6.4 is useful to follow for several subject matters.

FIGURE 6.4 Framework for crafting a written assessment of students' attitudes, knowledge, beliefs, and experiences about a topic.

	Area assessed	Example question
1.	Students' attitudes about the topic.	"I think meteorology is *boring*, *interesting*, etc."
2.	Students' school experiences with the topics.	"Have you ever studied meteorology or the weather? When?"
3.	Students' knowledge of an explanatory model centrally important in the unit.	"Explain what makes it rain. Include a diagram if you wish."
4.	Students' awareness of common knowledge associated with the topic.	"Imagine you are a TV or radio weather announcer. Write a forecast for what the weather will be tomorrow."
5.	Students' knowledge of technical terms associated with the topic.	"Describe what each of these instruments does or is used for: barometer, thermometer, and weather vane."
6.	Students' personal experiences with some aspect of the topic.	"Describe your most unusual or scary experience involving weather."

Source: Adapted from "Instructional Assessments: Lever for Systematic Change in Science Education Classrooms," by B. Gong, R. Venezky, & D. Mioduser, 1992, *Journal of Science Education and Technology, 1*(3), pp. 164–165. With kind permission of Springer Science and Business Media and the author.

Pretesting for Metacognition Skills

Some teachers have found it useful to pretest students' abilities to monitor and control their own thinking as they perform learning activities (Tittle, 1989; Tittle, Hecht, & Moore, 1993). If students are aware that learning one thing is more difficult than another, if they are able habitually to check statements before accepting them as facts, or if they habitually plan their work before beginning it, they are using *metacognitive skills*. You may wish to assess these skills before teaching so you will have a better idea of how well your students can monitor and control their thinking about the assignments you will make during the unit. You may wish to integrate teaching some of the metacognitive skills into the unit. The details for doing this are presented in Appendix F.

DIFFERENTIATING INSTRUCTION

Differentiated instruction refers to instructional practices that are altered to meet the needs, abilities, interests, and motivations of students. Characterized by clearly focused learning goals, preassessment and responses, flexible grouping, appropriate student choice during instruction, and ongoing formative assessment, some would say differentiated instruction is simply good instruction—instruction that is responsive to students' needs in the context of the standard or content being taught.

Our point in mentioning differentiated instruction in this chapter is that assessment planning is required to support it. Differentiated instruction relies on accurate, timely assessment in order to be effective. Appropriate preassessment will help you find out about students' prior knowledge, readiness, and interests regarding the learning target. Ongoing formative assessment will keep students engaged and in charge of regulating their thoughts and actions during instruction. Ongoing formative assessment will also help you make instructional decisions and help you decide when it's time for summative assessment.

In the example in Figure 6.3, assessment information most pertinent to differentiated instruction will be obtained from the pretest, observation and oral questioning, and homework—that is, toward the more formative end of the planning spectrum. Results from quizzes and independent investigations will also help. Differentiated instruction does not mean you never do whole-group activities. It does mean that you constantly review assessment information to maximize the effectiveness of the particular grouping, instructional activity, or assignment you decide on for each student, in order to get students ready for both individualized and undifferentiated work (Wormeli, 2006). To support these purposes, all your assessments, but especially your formative assessments, must be deeply aligned with your learning goals, must be frequent, and must produce timely responses and instructional decisions. Differentiated instruction, in short, relies on high-quality assessment and valid educational decisions based on that assessment more intensely than undifferentiated instruction does. Assessment planning is crucial for this approach.

CRAFTING A PLAN FOR ONE SUMMATIVE ASSESSMENT

This section focuses on one assessment purpose: using the results to help assign grades to students. This is an important responsibility, and you should not base this action on only one test. Chapter 14 will discuss strategies and techniques for assigning grades. In this section, however, our focus is narrower—*how to develop a plan* for one formal assessment instrument you will use for this summative purpose.

Organizing a Blueprint

Before creating a test, make a **blueprint** to describe both the content the assessment should cover and the performance expected of the student in relation to that content. Some authors call the blueprint a **table of specifications**. The blueprint serves as a basis for setting the number of assessment tasks and for ensuring that the assessment will have the desired emphasis and balance. Thus, the **elements of a complete test plan** include (a) content topics to assess, (b) types of thinking skills to assess, (c) specific learning targets to assess, and (d) emphasis (number of item or points) for each learning target to be assessed. Figure 6.5 illustrates such a blueprint for a science unit on cells. See Appendix G for examples of alternate ways of constructing a test blueprint for this same assessment.

The row headings along the left margin list the major topics (the knowledge) the assessment will cover. You can use a more detailed outline if you wish. The column headings across the top list the first three major cognitive process classifications

FIGURE 6.5 **Example of a blueprint for summative assessment of a science unit.**

Content outline	Remember	Understand	Apply	Total
I. *Basic Parts of Cell* A. *Nucleus* B. *Cytoplasm* C. *Cell membrane*	*Names and tells functions of each part of cell* 3 points	*Labels parts of cell shown on a line drawing* 3 points	*Given photographs of actual plant and animal cells, labels the parts* 2 points	8 points 40%
II. *Plant vs. Animal cells* A. *Similarities* B. *Differences* 1. *cell wall vs. membrane* 2. *food manufacture*		*Describes the cell wall and cell membrane* *Explains differences between plant and animal cells* 2 points		2 points 10%
III. *Cell Membrane* A. *Living nature of* B. *Diffusion* C. *Substances diffused by cells*	*Lists substances diffused and not diffused by cell membranes* *Gives definition of diffusion* 3 points	*Distinguishes between diffusion and oxidation* 1 point		4 points 20%
IV. *Division of Cells* A. *Phases in division* B. *Chromosomes and DNA* C. *Plant vs. Animal cell division*	*Gives definitions of division, chromosomes, and DNA* *States differences between plant and animal cell division* 4 points		*Given the numbers of chromosomes in a cell before division, states the number in each cell after division* 2 points	6 points 30%
Total	10 points 50%	6 points 30%	4 points 20%	20 points 100%

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 26), by A. J. Nitko and T-C. Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of the authors.

of the revised Bloom taxonomy. This test does not tap the "analyze, evaluate, or create" levels of thinking. You might use a project or other performance assessment to address those. You may use one of the other taxonomies, described in more detail in Appendix D, if you prefer.

The body of the blueprint lists the specific learning targets. Both a content topic and a level of complexity of the taxonomic category thus doubly classify the learning targets. In this example, most of the learning targets are at the lower and middle levels of the taxonomy. For a different emphasis, you would use the blueprint to identify the cells in which to write other objectives to assess.

Blueprint Specifies Assessment Emphasis

The numbers in the blueprint in Figure 6.5 describe the emphasis of the assessment, both in terms of percentage of the total number of tasks and in terms of the percentage of tasks within each row or content category. You decide how many tasks to include on an assessment after you consider (a) the importance of each learning target, (b) type of tasks, (c) content to be assessed, (d) what you emphasized in your teaching, and (e) amount of time available for assessment.

The example in Figure 6.5 illustrates how the blueprint can be useful as a planning tool. Summing across rows, you can see what portion of the test measures each content category. If the weight is not as intended, adjust the blueprint. Adjusting the blueprint *before* you write test items is much more efficient than editing a whole test after it is written. Summing down columns, you can see what portion of the test taps different cognitive levels, as well. Again, if the weight is not as intended, adjust the blueprint, allocating the points where they need to be to match your learning targets. Then (and only then), write the test.

The kind of blueprint in this example works with an objective test, where each item is worth 1 point, as well as for tests with multipoint items (for example, essay questions or problems to solve). Three points from the blueprint could be three 1-point items, or one 3-point item, or any combination of items worth 3 points.

Students will expect the various numbers of points on the assessment to correspond to the amount of time devoted to the material in class and to the emphasis they perceive you have placed on that material. If the assessment you are planning does not meet this expectation, it seems fair to notify the students of this fact well in advance of administering it.

This advanced planning for developing a summative classroom assessment allows you to view the assessment as a whole. In this way, you can maintain whatever balance or emphasis of content coverage and whatever complexity of performance you believe is necessary to match your teaching, and the assessment will be neither too easy nor too hard for your students. Plus, it simplifies the task of writing the test: It is easier to do that when a blueprint tells you exactly what kind of tasks and items you need. In addition, the blueprint is an excellent way to ensure that many of the criteria for improving validity (see Chapter 3) are met. Figure G.1 in Appendix G is a checklist for judging the quality of your blueprint.

Craft Blueprints Over Time

You need not attempt to devise a formal plan for all units in one semester or year. If you develop a blueprint for a few units each year, after a few years most units will have blueprints. As the learning targets change, you can update these blueprints with less work than originally required. Also, several teachers could draft blueprints for different units in a subject and exchange them. Even if a colleague's blueprint has to be modified to suit your particular teaching approach, you will likely save considerable time. When changes in the blueprints do occur, you should revise and redistribute the blueprints.

Simplified Specifications

It is difficult to classify learning objectives and items into the categories of some of the taxonomies. This approach to assessment planning is still useful, however. The purpose of formally laying out this two-way grid is not to promote exact or rigorous classification. Rather, it is a tool to help you recall the higher-order cognitive skills that need to be systematically taught and evaluated in the classroom. Sometimes, teachers merge categories (see examples in Appendix G) to simplify planning and yet maintain the blueprint's purpose of accurately representing desired cognitive levels.

Accommodations to the Summative Assessment

Any modifications you have made to the items on the test, conditions of administration, or student response modes in order to accommodate students with disabilities must give you assessment information that is valid. Review the discussion of accommodations in Chapter 5 and the list in Figure 5.2. Make sure that your accommodations for students who have IEPs are consistent with these educational plans. For example, if you are using a modified set of learning objectives for a student, consistent with the IEP, then a modified blueprint should be used to ensure the test reflects those learning objectives. If a student's IEP specifies that test items should be read to her, then (unless the test is a reading test) you should plan the logistics (the reader, a quiet location, etc.) to allow this to happen.

Blueprints for Student-Centered Assessment

As you can see, the assessment blueprint is a concise way to explain what is important for students to learn and to decide how much to emphasize each learning target in students' summative evaluations. Blueprints are useful instructional tools, too, especially with students in middle and senior high school. Therefore, share your assessment blueprints with your students.

Ideally, you should do this sharing when you begin the unit. You should review and discuss the blueprint thoroughly with the students to ensure they (a) have no misunderstandings, (b) understand the unit's emphasis, (c) understand what they will be held accountable for performing, and (d) see how the summative assessment factors into their overall grades. In Chapter 5 we discussed your professional responsibility to give students sufficient information when administering assessments. A blueprint is an excellent way to provide this information to middle and senior high school students. Older students may offer suggestions for changing the emphasis or manner of assessment, thus more fully engaging in their own learning and evaluation. Students can write test questions for each blueprint cell. Use them for a practice test.

CRITERIA FOR IMPROVING THE VALIDITY OF ASSESSMENT PLANS

In this section, we emphasize some of the general **criteria for evaluating a planned assessment**. These include (a) matching tasks to learning targets, (b) covering important skills, (c) selecting appropriate assessment task formats, (d) making assessments understandable, (e) satisfying validity criteria, (f) satisfying reliability criteria, (g) ensuring equivalence, and (h) identifying appropriate complexity and difficulty of tasks.

Two important validity principles from Chapter 3 should guide your assessment plans. Keep these principles at the forefront of your thinking, whether planning for a single assessment or multiple assessments:

1. Assessment results are valid only for specific interpretations and uses, not for all interpretations or uses.
2. Because no single assessment method gives perfectly valid results, more than one method should be used to assess the same achievement.

Figure 6.6 summarizes the main criteria and ways to improve the validity of your classroom assessments.

Matching Assessment Tasks to Learning Targets

Your teaching is most effective when your lesson plans, teaching activities, and learning targets are all aligned. All three should also be aligned with your state's curriculum framework and standards (see Chapter 2). Your assessment plans specify the important learning targets to be taught and assessed.

It is most important, therefore, that the actual tasks students perform on the assessment match those learning targets. To be valid, assessment procedures must match the learning targets. For example, if a learning target calls for students to build a model, write a poem, collect data, or perform a physical skill, the students should be administered a performance assessment. If your assessment task requires students only to list the parts of a model, to analyze an existing poem, to summarize data already collected, or to describe the sequence of steps needed for performing the physical skill, it does not match these learning targets. The validity of your classroom assessment results plummets when even *some* of the tasks do not match the stated learning targets.

As an example, consider the ninth-grade social studies learning target stated here and the three assessment tasks that follow it:

Example

Learning target: Students will explain in their own words the meaning of the concept of *culture*.

Task 1. Name three things that are important to the *culture* of indigenous Americans.

Task 2. Give a short talk to the class comparing three different *cultures*. In your talk, make sure you describe the similarities and differences among the cultures you have chosen.

Task 3. Write a paragraph telling in your own words what is meant by the term *culture*.

Only Task 3 matches the stated learning target. Consider Task 1: The performance required applies to a specific cultural situation rather than to the general concept of culture as intended by the learning target. Task 1 should not be used for the assessment of this learning target. The performances required in Task 2 are to "compare" and to "describe" ("giving a talk" is only the way the student has to indicate she is describing similarities and differences among the cultures). Although these are worthwhile activities, they seem to go beyond the more limited scope and main intent of the learning target. Because this task fails to match

FIGURE 6.6 **Criteria and ways to improve the validity of your classroom assessments.**

Criteria to use	Ways to evaluate your assessment plan
Align assessment tasks with curriculum, standards, and instruction	■ Be sure you clearly understand the main intent of the learning target and instruction to be taught and assessed. ■ *Think:* What is the main intent of the learning target? Does the assessment task require a student to do exactly as the main intent requires? ■ Analyze the assessment task to identify which part(s) may not match the learning target. Eliminate or rewrite the nonmatching parts.
Assess only important learning targets	■ Review the learning targets taught and assessed; prioritize them from most to least important. Eliminate assessments matching low-priority learning targets. ■ Be sure your state's standards or learning expectations are assessed by one or more of your assessments. ■ Create assessments that require students to demonstrate more than one high-priority learning target through the same task.
Use appropriate multiple assessment formats	■ Become skilled in creating many types of assessment formats. ■ Learn the strengths and limitations of each type of assessment format. ■ Analyze each learning target. *Think:* What are several different ways I can assess this achievement? How can I use two or more ways? ■ Analyze the assessment tasks to identify which parts may not match the learning target. Eliminate or rewrite the nonmatching parts. ■ Plan for assessing each important learning target in two or more ways.
Make assessments understandable	■ Be sure each assessment procedure has clear directions to the student, and that you have prepared students concerning each assessment. ■ Learn to craft assessments well so they will satisfy the criteria and checklists contained in Chapters 8 through 13. ■ Learn to craft scoring rubrics well so they will satisfy the criteria and checklists in Chapter 12.
Follow appropriate validity criteria	■ Use the criteria described in Chapter 3.
Use appropriate length of assessments	■ Be sure all students who know the material can finish within the time limits. ■ Follow the suggestions for improving reliability given in Chapter 4.
Ensure equivalence across years	■ Use blueprints from previous years to guide you in crafting this year's assessment blueprints. ■ Be sure to make the difficulty and complexity of this year's assessment tasks equivalent to last year's tasks.
Ensure appropriate difficulty and complexity of assessment tasks	■ Be sure the conditions and tools for students to use during the assessment are appropriate for the learning targets and the students' educational development. ■ Add appropriate accommodations for students with disabilities (see Figure 5.2).

the learning target, it should not appear on the assessment either.

What should you do if you create or identify a "great" task that does not match the stated learning target? You have only three choices: disregard the task, modify the task so it matches the learning target, or modify the learning target so it matches the task. Often, crafting an excellent assessment task helps further clarify a learning target: We see the full meaning of the learning target, which was not previously clear from its verbal statement. If this is the case, then you should modify the stated learning target so it more clearly expresses what you intend.

Be careful, however. If you have already communicated the assessment plan to students, you need to be sure that you do not "surprise" them with a more complex or difficult task than the type for which they are preparing themselves. Changing the rules in midstream is often unethical. Usually, it guarantees that you lose the respect of at least some students. Rather than completely discarding the task, you could either modify it to suit the stated learning target or save it until the next time you teach the unit. At that time you can more clearly specify the learning target.

Using Appropriate Multiple Assessment Task Formats

Many varieties of assessments are available. One of the criteria we discussed in Chapter 3 (Figure 3.1) is to present students with multiple ways to demonstrate their competence. The validity of your assessment results usually improves, therefore, if you use several **task formats** (paper-and-pencil tests, performance assessments, and personal communication) to assess students. Try to use the combination of formats that *most directly assesses the intents* of the stated learning targets.

Making Assessments Understandable to Students

As you plan and create your assessments, remember that you need to make clear to students how and when they will be assessed, what they will be required to do, and when and how they will be evaluated. A section in Chapter 13, “Preparing Students for Assessment,” describes some of the information about your planned assessment that your students need to know. In addition, you will want to be sure the directions to students, the assessment tasks (e.g., your test questions), and the scoring rubrics (e.g., criteria for full marks) are understandable to all students. For example, according to Jakwerth, Stancavage, and Reed (1999), students who leave constructed-response questions unanswered on the National Assessment of Educational Progress may do so because they “couldn’t figure out what the question was asking” (p. 9), “didn’t really get the question” (p. 9), “thought it would take too long” (p. 9), or “didn’t realize [I] had to do both parts” (p. 10). You can avoid such problems by clearly writing and explaining your assessments to students and, in turn, being sure that students understand before they begin.

Satisfying Appropriate Validity Criteria

The main criteria for judging the quality of your assessment are validity criteria. In Chapter 3 (Figure 3.1), we discussed seven categories of validity criteria for classroom assessments: (1) content representativeness and relevance; (2) thinking processes and skills represented; (3) consistency with other assessments; (4) reliability and objectivity; (5) fairness to different types of students; (6) economy, efficiency, practicality, and instructional features; and (7) multiple assessment usage. We cannot overstate how critical these validity criteria are for effective assessment.

Satisfying Appropriate Reliability Criteria

In Chapter 4, we discussed reliability concerns for classroom assessment (see Figure 4.1). The length of your assessment is one of the factors affecting reliability. Length depends on three major factors: (1) the amount of time you have available for assessment, (2) the students’ educational development, and (3) the level of reliability you wish the results to have. Longer assessments are more reliable than shorter assessments. Classroom assessments should be power assessments: That is, every student who has learned the material should have enough time to perform each task. Your experience with the subject matter and the students you teach will help you decide how long to make the assessment.

As practical guidelines, use the time suggestions in Figure 6.7 for students in middle and senior high school. In 40 minutes of assessment, for example, you can administer a test with a short essay and 15 to 20 complex multiple-choice items. Modify these time suggestions to suit your students as your experience deepens.

Remember, too, that students will be taking state-mandated and other standardized tests: These tests are typically 40 to 60 minutes in length,

FIGURE 6.7 **Time requirements for certain assessment tasks.**

Type of task	Approximate time per task (item)
True-false items	20–30 seconds
Multiple-choice (factual)	40–60 seconds
One-word fill-in	40–60 seconds
Multiple-choice (complex)	70–90 seconds
Matching (5 stems/6 choices)	2–4 minutes
Short-answer	2–4 minutes
Multiple-choice (w/calculations)	2–5 minutes
Word problems (simple arithmetic)	5–10 minutes
Short essays	15–20 minutes
Data analyses/graphing	15–25 minutes
Drawing models/labeling	20–30 minutes
Extended essays	35–50 minutes

even for elementary students. Your classroom assessments, therefore, should give students the opportunity to practice taking longer assessments. You do not want the mandated assessment to be the first long test students take each year.

Ensuring Equivalence

If the content of the units you are assessing has remained essentially the same since the last time you taught them, your summative assessment instruments on the two occasions should be equivalent. Building this semester's assessment instruments to last semester's blueprints increases the likelihood the two instruments will be equivalent, even if you use different questions. Blueprints will help ensure that both years' assessments cover the same content and thinking skills and emphasize the same knowledge and skills. Equivalent instruments are fairer to students. **Equivalence** means that students past and present are required to know and perform tasks of similar complexity and difficulty to earn the same grade. Of course, if you changed the content or learning targets of the unit, the blueprints and the assessment should change as well. Also, if results of your past assessments were unsatisfactory, you should not perpetuate them from year to year.

WHAT RANGE OF ASSESSMENT OPTIONS IS AVAILABLE?

Whether you are assessing for summative or formative purposes, you have a wide range of options at your disposal. Which should you use? Before deciding, you need to know three things: (1) the learning targets students should achieve, (2) the purpose for which you want to use the assessment results, and (3) the advantages of an assessment technique for the specific purpose you have in mind. This section discusses the general advantages and disadvantages of the many assessment options available to you. The assessment you use should be the most appropriate for assessing the learning targets you wish students to achieve. You should defend your choice(s) on the basis of the validity and reliability of the results.

The most commonly used types of classroom assessment procedures are listed in Figure 6.8 along with their advantages, their limitations, and brief suggestions for improved use. The techniques are grouped into two categories: formative assessment techniques and summative assessment techniques.

Formative Assessment Options

Formative assessments gather information to help improve students' achievement of learning targets. This information guides and fine-tunes both your thinking and your students'. You use formative assessment information to plan your next teaching activities, to diagnose the causes of students' learning difficulties, and to give students information about how to improve. In fact, assessment is not truly "formative" unless students actually use the information for improvement.

You gather formative information while you are still teaching the material and while students are still learning it. As a result, these are often **informal assessment techniques**. That is, they occur spontaneously as you need information, and you rarely stop teaching to do the assessment. Figure 6.8 summarizes eight categories of formative assessment options. The eight categories fall into three groups as described in the following paragraphs.

Oral Assessment Techniques

You may gather information to improve students' learning without creating tests or other paper-and-pencil tasks. Conversations with teachers who have taught a student may give you insight into the student's background and which approaches have worked in the past. These conversations may also help as you size up the class at the beginning of the term. Conversations with students give you additional insight into their feelings, attitudes, interests, and motivations.

As you teach a lesson, you question students about the material. These questions should encourage students to think about the material and to reveal their understandings, including misconceptions. This will help you guide your teaching. Avoid the "recitation" type of questioning in which you seek short answers to your questions. This style of questioning provides little insight into students' thinking and, therefore, provides little formative information. Avoid the tendency to ignore or ask only simple questions of the shy and less verbal students (Good & Brophy, 2002). Consider using whiteboards, letter cards, or hand signals for student responses so you can survey all students' answers, not just a few, for each question.

FIGURE 6.8 **Advantages, limitations, and pitfalls of alternative types of classroom assessment techniques.**

Assessment alternatives	Advantages for teachers	Disadvantages for teachers	Suggestions for improved use
		Formative assessment techniques	
1. Conversations and comments from other teachers	(a) Fast way to obtain certain types of background information about a student. (b) Permit colleagues to share experiences with specific students in other learning contexts, thereby broadening the perspective about the learners. (c) Permit attainment of information about a student's family, siblings, or peer problems that may be affecting the student's learning.	(a) Tend to reinforce stereotypes and biases toward a family or social class. (b) Students' learning under another teacher or in another context may be quite unlike their learning in the current context. (c) Others' opinions are not objective, often based on incomplete information, personal life view, or personal theory of personality.	(a) Do not believe hearsay, rumors, biases of others. (b) Do not gossip or reveal private and confidential information about students. (c) Keep the conversation on a professional level and focused on facts.
2. Casual conversations with students	(a) Provide relaxed, informal setting for obtaining information. (b) Students may reveal their attitudes and motivations toward learning that are not exhibited in class.	(a) A student's mind may not be focused on the learning target being assessed. (b) Inadequate sampling of students' knowledge; too few students assessed. (c) Inefficient students' conversation may be irrelevant to assessing their achievement.	(a) Do not appear as an inquisitor, always probing students. (b) Be careful not to misperceive a student's attitude or a student's degree of understanding.
3. Questioning students during instruction	(a) Permits judgments about students' thinking and learning progress during the course of teaching; gives teachers immediate feedback. (b) Permits teachers to ask questions requiring higher-order thinking and elaborated responses. (c) Permits student-to-student interaction to be assessed. (d) Permits assessment of students' ability to discuss issues with others orally and in some depth.	(a) Some students cannot express themselves well in front of other students. (b) Requires education in how to ask proper questions and to plan for asking specific types of questions during the lesson. (c) Information obtained tends to be only a small sample of the learning outcomes and of the students in the class. (d) Some learning targets cannot be assessed by spontaneous and short oral responses; they require longer time frames in which students are free to think, create, and respond. (e) Records of students' responses are kept only in the teacher's mind, which may be unreliable.	(a) Be sure to ask questions of students who are reticent or slow to respond. Avoid focusing on verbally aggressive "stars." (b) Wait 5–10 seconds for a student to respond before moving on to another. (c) Avoid limiting questions to those requiring facts or a definite correct answer, thereby narrowing the focus of the assessment inappropriately. (d) Do not punish students for failing to participate in class question sessions or inappropriately reward those verbally aggressive students who participate fully. (e) Remember that students' verbal and nonverbal behavior in class may not indicate their true attitudes/values.
4. Daily homework and seatwork	(a) Provide formative information about how learning is progressing. (b) Allow errors to be diagnosed and corrected. (c) Combine practice, reinforcement, and assessment.	(a) Tend to focus on narrow segments of learning rather than integrating large complexes of skills and knowledge. (b) Sample only a small variety of content and skills on any one assignment.	(a) This method assesses learning that is only in the formative stages. It may be inappropriate to assign summative letter grades from the results. (b) Failure to complete homework or completing it late is no reason to punish students by

		(c) Assignment may not be complete or may be copied from others.	embarrassing them in front of others or by lowering their overall grade. Learning may be subsequently demonstrated through other assessments. (c) Do not inappropriately attribute poor test performance to the student not doing the homework. (d) Do not overemphasize the homework grade and overuse homework as a teaching strategy (e.g., using it as a primary teaching method.)
5. Teacher-made quizzes and tests	(a) Although primarily useful for summative evaluation, they may permit diagnosis of errors and faulty thinking. (b) Provide for students' written expression of knowledge.	(a) Require time to craft good tasks useful for diagnosis. (b) Focus exclusively on cognitive learning targets.	(a) Do not overemphasize lower-level thinking skills. (b) Use open-ended or constructed-response tasks to gain insight into a student's thinking processes and errors. (c) For better diagnosis of a student's thinking, use tasks that require students to apply and use their knowledge to "real-life" situations.
6. In-depth interviews of individual students	(a) Permit in-depth probing of students' understandings, thinking patterns, and problem-solving strategies. (b) Permit follow-up questions tailored to a student's responses and allow a student to elaborate answers. (c) Permit diagnosis of faulty thinking and errors in performances.	(a) Require a lot of time to complete. (b) Require keeping the rest of the class occupied while one student is being interviewed. (c) Require learning skills in effective educational achievement interviewing and diagnosis.	(a) If assessing students' thinking patterns, problem-solving strategies, etc., avoid prompting student toward a prescribed way of problem solving. (b) Some students need their self-confidence bolstered before they feel comfortable revealing their mistakes.
7. Growth portfolios	(a) Allow large segments of a student's learning experiences to be reviewed. (b) Allow monitoring a student's growth and progress. (c) Communicate to students that growth and progress are more important than test results. (d) Allow student to participate in selecting and evaluating material to include in the portfolio. (e) Can become a focus of teaching and learning.	(a) Require a long time to accumulate evidence of growth and progress. (b) Require special effort to teach students how to use appropriate and realistic self-assessment techniques. (c) Require high-level knowledge of the subject matter to diagnose and guide students. (d) Require the ability to recognize complex and subtle pattern of growth and progress in the subjects. (e) Results tend to be inconsistent from teacher to teacher.	(a) Be very clear about the learning targets toward which you are monitoring progress. (b) Use a conceptual framework or learning progress model to guide your diagnosis and monitoring. (c) Coordinate portfolio development and assessment with other teachers. (d) Develop scoring rubrics to define standards and maintain consistency.

(*continued*)

FIGURE 6.8 (*Continued*)

Assessment alternatives	Advantages for teachers	Formative assessment techniques	Suggestions for improved use
8. Attitude and values questionnaires	(a) Assess affective characteristics of students. (b) Knowing student's attitudes and values in relation to a specific topic or subject matter may be useful in planning teaching. (c) May provide insights into students' motivations.	(a) The results are sensitive to the way questions are worded. Students may misinterpret, not understand, or react differently than the assessor intended. (b) Can be easily "faked" by older and testwise students.	(a) Remember that the way questions are worded significantly affects how students respond. (b) Remember that attitude questionnaire responses may change drastically from one occasion or context to another. (c) Remember that your personal theory of personality or personal value system may lead to incorrect interpretations of students' responses.
Summative assessment techniques			
1. Teacher-made tests and quizzes	(a) Can assess a wide range of content and cognitive skills. (b) Can be aligned with what was actually taught. (c) Use a variety of task formats. (d) Allow for assessment of written expression.	(a) Difficult to assess complex skills or ability to use combinations of skills. (b) Require time to create, edit, and produce good items. (c) Class period is often too short for a complete assessment. (d) Focus exclusively on cognitive outcomes.	(a) Do not overemphasize lower-level thinking skills. (b) Do not overuse short-answer and response-choice items. (c) Craft task requiring students to apply knowledge to "real life."
2. Task focusing on procedures and processes	(a) Allow assessments of nonverbal as well as verbal responses. (b) Allow students to integrate several simple skills and knowledge to perform a complex, realistic task. (c) Allow for group and cooperative performance and assessment. (d) Allow assessment of steps used to complete an assignment.	(a) Focus on a narrow range of content knowledge and cognitive skills. (b) Require a great deal of time to properly formulate, administer, and rate. (c) May have low inter-rater reliability unless scoring rubrics are used. (d) Students' performance quality is not easily generalized across different content and tasks. (e) Tasks that students perceive as uninteresting, boring, or irrelevant do not elicit the students' best efforts.	(a) Investigate carefully the reason for student's failure to complete the task successfully. (b) Use a scoring rubric to increase the reliability and validity of results. (c) Do not confuse the evaluation of the process a student uses with the need to evaluate the correctness of the answers. (d) Allow sufficient time for students to adequately demonstrate the performance.
3. Projects and tasks focusing on products	(a) Same as 2(a), (b), and (c).	(a) Same as 2(a), (b), (c), (d), and (e).	(a) Same as 2(a), (b), (c), and (d).

	(b) Permit several equally valid processes to be used to produce the product or complete the project. (c) Allow assessment of the quality of the product. (d) Allow longer time than class period to complete the tasks.	(b) Students may have unauthorized help outside class to complete the product or project. (c) All students in the class must have the same opportunity to use all appropriate materials and tools in order for the assessment to be fair.	(b) Give adequate instruction to students on the criteria that will be used to evaluate their work, the standards that will be applied, and how students can use these criteria and standards to monitor their own progress in completing the work. (c) Do not mistake the aesthetic appearance of the product for substance and thoughtfulness. (d) Do not punish tardiness in completing the project or product by lowering the student's grade.
4. Best works portfolios	(a) Allow large segments of a student's learning experience to be assessed. (b) May allow students to participate in the selection of the material to be included in the portfolio. (c) Allow either quantitative or qualitative assessment of the works in the portfolio. (d) Permit a much broader assessment of learning targets than tests.	(a) Require waiting a long time before reporting assessment results. (b) Students must be taught how to select work to include as well as how to present it effectively. (c) Teachers must learn to use a scoring rubric that assesses a wide variety of pieces of work. (d) Inter-rater reliability is low from teacher to teacher. (e) Require high levels of subject-matter knowledge to evaluate students' work properly.	(a) Be very clear about the learning targets to be assessed to avoid confusion and invalid portfolio assessment results. (b) Teach a student to use appropriate criteria to choose the work to include. (c) Do not collect too much material to evaluate. (d) Coordinate portfolio development with other teachers. (e) Develop and use scoring rubrics to define standards and maintain consistency.
5. Textbook-supplied tests and quizzes	(a) Allow for assessment of written expression. (b) Already prepared, save teachers' time. (c) Match the content and sequence of the textbook or curricular materials.	(a) Often do not assess complex skills or ability to use combinations of skills. (b) Often do not match the emphases and presentations in class. (c) Focus on cognitive skills. (d) Class period is often too short for a complete assessment.	(a) Be skeptical that the items were made by professionals and are of high quality. (b) Carefully edit or rewrite the item to match what you have taught. (c) Remember that you are personally responsible for using a poor-quality test. You must not appeal to the authority of the textbook.
6. Standardized achievement tests	(a) Assess a wide range of cognitive abilities and skills that cover a year's learning. (b) Assess content and skills common to many schools across the country.	(a) Focus exclusively on cognitive outcomes. (b) Often the emphasis on a particular test is different from the emphasis of a particular teacher.	(a) Avoid narrowing your instruction to prepare students for these tests when administrators put pressure on teachers. (b) Do not use these tests to evaluate teachers.

(*continued*)

FIGURE 6.8 (*Continued*)

Assessment alternatives	Advantages for teachers	Formative assessment techniques	Suggestions for improved use
	(c) Items developed and screened by professionals, resulting in only the best items being included. (d) Corroborate what teachers know about pupils; sometimes indicate unexpected results for specific students. (e) Provide norm-referenced information that permits evaluation of students' progress in relation to students nationwide. (f) Provide legitimate comparisons of a student's achievement in two or more curricular areas. (g) Provide growth scales so students' long-term educational development can be monitored. (h) Useful for curriculum evaluation.	(c) Do not provide diagnostic information. (d) Results usually take too long to get back to teachers, so are not directly useful for instructional planning.	(c) Do not confuse the quality of the learning that did occur in the classroom with the results on standardized tests when interpreting them. (d) Educate parents about the tests, limited validity for assessing a student's learning potentials.

A good way to plan your oral questioning is to use a thinking skills taxonomy. In every lesson, be sure you ask several questions from the higher-order thinking categories of the taxonomy. Below are examples of some questions a teacher might ask students who have been studying the short story as a literary form:

Example

Remember	"Who was the main character in the last story we read?"
Understand	"What were some of the personal problems that the characters in this story had to solve?"
Apply	"Are the characters' problems in any way similar to the problems you or someone you know have had? Tell us about that. Don't use real names if you will embarrass the person."
Analyze	"What literary devices, style of writing, or 'writing trick' did the author use to help the reader really understand how the characters were feeling? Explain how this was done."
Evaluate	"What are three or four criteria that we can apply to all of the stories so we can compare and evaluate their literary quality?"
Create	"So far this semester, we have read eight short stories. In each one, a character (sometimes two characters) wasn't able to solve his or her problem satisfactorily—even though each character tried to do so. Why is that? What do they all have in common that resulted in failure to solve their problems? What general problem-solving approach did all of these characters use that resulted in their failure?"

Paper-and-Pencil Assessment Techniques Each day you give students seatwork and homework. These **paper-and-pencil assessments** let students practice the learning targets and perhaps extend their learning beyond the specific material you taught. You should review the results of seatwork and homework not just for correctness, but for whether the work reveals students' errors or faulty thinking that needs correction. If a student is exhibiting a pattern of errors, the student may have a misconception or may be using a rule consistently but inappropriately. (See Chapter 7 for an example.) Providing that specific information as feedback to students who need it is a powerful way of personalizing learning and helps students change.

You also periodically create and administer short quizzes and tests. These monitor the progress students are making toward achieving learning targets. Tests and exams tend to be somewhat formal and are more useful for summative evaluation of students than for formative evaluation. However, if you use open-ended response items and carefully review students' responses for insights into their thinking, you will be able to derive some diagnostic information from these techniques.

Portfolios Other formative evaluation techniques are somewhat more labor intensive than the ones we have discussed so far. A **growth portfolio** is a selected sequence of a student's work that demonstrates progress or development toward achieving the learning target(s). By containing "not-so-good works," "improved works," and "best works," it shows progress and learning during the course.

Typically, both the teacher and the student decide what a portfolio should include. Further, students are usually asked to describe the work they included, why they selected it, what it demonstrates about their learning, and their affective reactions to the material and to their learning experiences. Because a portfolio is built up over time, it permits closer integration of assessments with instruction than with some of the other techniques. These attributes are considered advantages of portfolios over one-shot assessment techniques because of the richness of information they provide the teacher.

Growth portfolios are usually evaluated qualitatively, although rating scales are sometimes used. Evaluating the evidence qualitatively requires a significant amount of skill and knowledge about student learning and the subject matter. The following excerpt from an evaluation of the language arts portfolio of an eighth-grade student illustrates both the richness of the information in the portfolio and the deep level of teacher knowledge required to evaluate it:

Example

Our experience is that growth is often manifested in qualitative changes in the writing—changes in the complexity of the problems that students undertake,

> which may involve losing control over other features of the writing like organization or mechanics. Take Gretchen . . . who included two pieces of expository response to literature in her portfolio. In one sense, the second piece is not as strong as the first—it is not well organized or coherent—but it is a richer interpretation. Unlike the first piece, which simply compares two groups of characters from *Lord of the Flies* . . . the second piece, on *Animal Farm*, has a thematic framework about the role of scapegoats that is played out with evidence from Gretchen's own personal experience, from the novel, and from a definition of the term acquired from another resource. A comparison of Gretchen's revisions in the two pieces shows a newly developed awareness of the need for elaboration and for evidence on particular points. (Moss et al., 1992, p. 13)

Chapter 12 describes portfolio construction and use in more detail.

Interviews In addition to portfolios, you may conduct *interviews with individual students*. Interviews can give you additional insights into students' thinking and learning difficulties. These interviews are more effective if you organize them around key concepts or specific problem-solving tasks. For example, you could work with the student to create a mental map of the relevant concepts in a unit and discuss with the student how he believes the concepts to be related to one another. Or, for example, many writing teachers use individual writing conferences with students based on drafts of written work. You may also administer a simple questionnaire to your class to gain insight into students' attitudes and values associated with the concepts you are about to teach. We saw a framework for this strategy in Figure 6.4.

Summative Assessment Options

Summative assessments help you formally evaluate students' learning-target achievement so you can report to students, parents, and school officials. This evaluation results in a home report or a report card grade. Summative assessment techniques are usually more formal than formative assessment techniques. Keep in mind, however, that formative and summative are not always distinct. For example, after you teach a unit, you may give a summative unit test. However, you may find students who have not achieved the learning targets. This will usually require you to reteach the students or provide remedial instruction. Because you have used the summative assessment to guide your teaching, it has provided formative assessment information.

Figure 6.8 shows six categories of summative assessment options. We may separate these into two groups: teacher-made assessments and external (extraclassroom) assessments.

Teacher-Made Assessments We have already mentioned tests and quizzes. These paper-and-pencil techniques may include open-ended questions (such as essays and other constructed-response formats), multiple-choice, true-false, and matching exercises. Chapters 8 through 11 discuss how to craft these formats.

But paper-and-pencil techniques are limited primarily to verbal expressions of knowledge. Students must read and respond to the assessment materials using some type of written response, ranging from simple marks and single words to complex and elaborated essays. Students' abilities to carry out actual experiments, to carry out library research, or to build a model, for example, are not assessed directly with paper-and-pencil techniques. Further, it is usually difficult for teachers to craft paper-and-pencil tasks that require students to apply knowledge and skills from several areas to solve real-life or "authentic" problems. Chapter 11 suggests techniques that assess higher-order thinking skills.

Performance assessment techniques require students to physically carry out a complex, extended **process** (e.g., present an argument orally, play a musical piece, or climb a knotted rope) or produce an important **product** (e.g., write a poem, report on an experiment, or create a painting). The performances you assess should (a) be very close to the ultimate learning targets, (b) require students to use combinations of many different abilities and skills, and (c) require students to perform under "realistic conditions" (especially requiring student self-pacing, self-motivation, and self-evaluation). Some performance assessments require paper-and-pencil as a medium for expression (e.g., writing a research paper or a short story), but the emphasis in these performances is on the complexity of the product, and students are allowed appropriate time limits. This distinguishes such performance assessments from the short answers, decontextualized math problems, or brief (one class period) essay tasks found on typical paper-and-pencil assessments.

Because some performance assessments very closely measure the ultimate learning targets of schooling, they may be used as instructional tools. For example, you may instruct a student on presenting arguments orally and require the student to perform the task several times over the course of the term. You might repeat the teaching-performance combination several times until the student has learned the technique to the degree of expertise appropriate to the student's level of educational development.

Principal disadvantages are that a great deal of time is required to craft appropriate tasks, to prepare marking schemes or rating scales, to carry out the assessment itself, and to administer several tasks. The last point relates to the validity of interpreting students' results. Seldom can you generalize a student's performance on one task to performance on another. That is, how well a student performs depends on the specific content and task to which the performance is linked (Baker, 1992; Linn, 1994). A student may write a good poem about the people in her neighborhood but an awful poem about the traffic in Los Angeles. How good is the student as a poet in such cases? Quality performance assessment requires a very clear vision of an important learning target and a high level of skill to translate that vision into appropriate tasks and grading criteria (Arter & Stiggins, 1992).

Previously, we discussed the growth portfolio as a formative assessment tool. Portfolios may also be used for summative evaluation. The **best works portfolio** is a representative selection of a student's best products (productions) that provides evidence of the degree to which the student has achieved specified learning targets. In an art course it might be the student's best works in drawing, painting, sculpture, craftwork, and, perhaps, a medium chosen by the student. In mathematics it might include reports on mathematical investigations, examples of how the student applied mathematics to a real problem, writings about mathematics or mathematicians, and examples of how to use mathematics in social studies, English, and science. Best works portfolios focus on summative evaluation. To improve reliability of portfolio evaluations, you need to craft a scoring rubric. Share the rubric with students and teach them how to select their best work in light of those rubrics. Chapter 12 presents more details on portfolio assessments.

External (Extraclassroom) Assessments Teachers often use two other techniques. One involves the *quizzes and tests supplied by textbook publishers*. These are convenient because you don't have to create them yourself, and they match the book you are using. The problem is that these assessment materials are often of *poor quality*: they may not match local learning targets very well, they tend to focus on low-level thinking skills, and they can be poorly crafted. As we mentioned in Chapter 5, you have a professional responsibility to improve these assessment materials before using them.

Standardized Achievement Test Standardized tests also provide summative assessment information. Unlike textbook tests, these materials are usually quite well crafted and supported by research on the validity of the scores. The tests consist of a battery of subtests, each covering a different curriculum area. Because the same group of students (norm group) took all subtests, the publisher's percentile norms allow you to compare a student's development in two or more curricular areas; and the publisher's score scales allow you to monitor a student's growth over time. Your own or your school district's tests cannot provide these types of information. A standardized test battery does not match your curriculum or your teaching goals exactly. Therefore, use it to assess broad goals (e.g., reading comprehension) rather than the specific learning targets in your classroom. You will learn more details about this assessment option by studying Chapters 15 through 18.

ASSESSMENT PLANNING FOR RESPONSE TO INTERVENTION

Based on a definition in the 1975 Education of All Handicapped Children Act, "underachievers" have historically been identified as students with IQ/achievement discrepancies: students whose classroom work does not reach the expectations for students of their ability. The 2004 Individuals with Disabilities Education and Improvement Act added a second definition by which such students could be identified: Students who do not progress in otherwise effective instruction are not responsive to that instruction (Fuchs & Fuchs, 2007; Klotz and Canter, 2006).

Response to Intervention (RTI) is therefore an initiative that many states are using not only to identify students in need of special assistance but also to provide tiers of assistance in order to minimize the number of students identified for special education services. A complete description of RTI is beyond the scope of this book, but we mention it here in the chapter on assessment planning because planning for and implementing RTI-related assessment is increasingly a part of teachers' work. The assessment principles you are learning in this book will help you with the assessment you do for RTI.

Assessment for RTI involves planning for screening and for progress monitoring. As students enter school in kindergarten, they are screened to see whether they are at risk for not responding to regular instruction, typically by assessing readiness for reading and mathematics. Students who are potentially at risk are identified for **progress monitoring**, which is regular, classroom-based assessment to assess how students are responding to classroom instruction.

Students who do not make progress—who are not responsive to the primary instruction—are identified for a first tier of assistance, typically some type of tutoring. Again progress monitoring charts their improvement (or not) on basic elements of classroom instruction (e.g., letter recognition, reading fluency, or math problem solving, depending on the student). A specified amount of increase in achievement on the classroom-based assessment identifies students who respond to this first-tier intervention. Students who do not make progress are further identified for a second tier of intervention, for example, more intensive tutoring. Progress monitoring continues, assessing the responsiveness of the student to the second tier of intervention. If students do not respond with increased achievement, then they are eligible for more comprehensive evaluation and potential third-tier intervention, for example, learning or behavioral disability certification and special education placement. Even then, progress monitoring continues.

The intent of such a system is that ongoing progress monitoring will help the intermediate interventions work. Many students, with additional assistance, can make progress and do not need the more acute intervention of special education placement. Progress monitoring is the means by which this is determined. Progress monitoring usually takes the form of what once was called curriculum-based measurement (CBM) or curriculum-based assessment (CBA) in the literature (Fuchs & Fuchs, 2007). These assessments are made up of tasks (like reading a passage) or items (like math problems) that are part of regular classroom instruction. Achievement is typically mapped by graphing scores, for example, number of words read correctly per minute or number of problems solved correctly in some fixed amount of time. Appropriate and responsive progress is typically defined as an increase in achievement (e.g., weekly) sufficient to reach the student's goal by some stated time, or by comparing the slope of the line to a cutoff level specified in a research-based program (Fuchs & Fuchs, 2007; Klotz & Canter, 2006).

If your school or district uses RTI methods, you will be involved in planning, administering, record keeping, and decision making for progress monitoring. Your understanding of assessment planning and assessment quality principles will help you with this work.

CONCLUSION

This chapter has introduced basic classroom assessment planning. The most important planning principle is to base both assessment and instruction on learning targets and a deep understanding of the essential knowledge and skills students need to achieve them. The planning principles in this chapter apply to both formative and summative assessment. Chapter 7 discusses diagnostic and formative assessment in more detail. This chapter has also introduced basic assessment formats. Chapters 8 through 12 show how to create assessments in each of these formats in turn.

EXERCISES

1. Visit a classroom (if you are not an in-service teacher) or use your own teaching experience to complete the following:
   a. Identify one or more specific examples of each classroom assessment purpose described in Figure 6.1.
   b. For each example, describe what assessment tools and information the teacher used to make that decision.
   c. Classify each tool or technique into one of the assessment-option categories shown in Figure 6.8.
2. Visit a classroom, or look around your own classroom, and list all the instructional resources that provide assessment or assessment-like tools.
   a. Classify each as true-false, multiple-choice, matching, essay, short-answer, completion, performance assessments, projects, portfolios, oral questioning strategies, observation strategies, or in-depth interviewing strategies.
   b. Which type(s) is (are) dominant?
   c. Tally the thinking skill levels each appears to assess. Which levels of thinking do the majority seem to assess?
   d. Judge the quality of each of these materials using the criteria in Figure 3.1.
3. Select a unit in your subject area for which you might craft a summative assessment instrument. Develop a complete blueprint for this assessment, using Figure 6.5 as a model. Describe the kinds of tasks you would include, and explain how you would decide whether the tasks matched the learning targets. Estimate the amount of time it would take students to complete your assessment.
4. Develop an assessment plan for a unit of instruction in your area. Using Figure 6.3 as a model, list lessons and learning targets, types of assessment, purpose(s) of assessment, and actions to take using assessment results. Share your results with your classmates.
5. Develop an assessment plan for a marking period or a semester in an area you teach. Using Figure 6.2 as a model, include the time frame for the units, the formative and summative assessment strategies, and the weighting of the assessments within units and across units (i.e., for the entire time periods). Share your results with your classmates

Diagnostic and Formative Assessments

KEY CONCEPTS

1. Diagnostic assessment is conducted to identify what knowledge and skills a student has mastered and potential reasons for nonmastery.
2. Six approaches to diagnostic assessment are each based on a different definition of a learning "deficit."
3. Formative assessment is a loop: Students and teachers focus on a learning target, evaluate current student work against the target, act to move the work closer to the target, and repeat the process. Unlike diagnostic assessment, formative assessment seeks to identify both strengths and weaknesses.
4. Cognitive benefits of formative assessment include providing the information students need in order to improve and giving the student practice at "learning how to learn." Motivational benefits of formative assessment include helping students feel in control of their own learning and supporting self-regulation.
5. A system in which good assignments, formative feedback and self-assessment, summative assessments, and scoring criteria all match the learning targets supports student learning.
6. A teacher can obtain formative assessment information by talking with students, observing them working, or looking at the work itself.
7. Learning progressions are developmental sequences that describe typical progress in understanding particular content or advancing a particular skill.
8. Formative assessment information for the student comes mainly in the form of feedback. Good feedback is descriptive, specific, and contains information for improvement.
9. It is important to record the results of formative assessments, look for patterns, and share your insights with students.

IMPORTANT TERMS

algorithmic knowledge

component competencies of problem-solving

concept mapping

deficits in learning

identifying errors in performance

knowledge structure

learning hierarchy

learning progression

linguistic knowledge

mastery of specific objectives

passing score

prerequisite knowledge and skill deficits

profiling content areas strengths and weaknesses

schematic knowledge

strategic knowledge

student self-assessment

surface feature

From Chapter 7 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

Some authors consider diagnostic assessment and formative assessment as two separate practices, one that takes place before instruction and one during instruction. Our perspective is that the older term diagnostic assessment and the newer term formative assessment are getting at a similar idea. Diagnostic assessment puts the emphasis on the teacher understanding the status of student learning for the sake of planning instruction. Formative assessment puts the emphasis on the students, as well as teachers, understanding the status of learning, for the purpose of identifying next steps to take for improvement (Assessment Reform Group, 2002). Consequently, a timing distinction does not seem completely useful. In this chapter, we present six approaches to diagnosing learning difficulties, which teachers will find useful for ferreting out problems in enough detail to address them specifically in lesson plans. Next, we describe more informal approaches, formative assessment strategies that can be implemented during the course of instruction and involve students. We hasten to add that "informal" does not mean "unplanned."

DIAGNOSTIC ASSESSMENT

Diagnostic assessment of learning difficulties serves two related purposes: (1) to identify which learning targets a student has not mastered and (2) to suggest possible causes or reasons why the student has not mastered the learning targets. The emphasis is on learning deficits, that is, remediation of what the student does not know.

SIX APPROACHES TO DIAGNOSIS OF LEARNING PROBLEMS

Different approaches to diagnosis provide different levels of detail about students' **deficits in learning**. They also differ in the degree to which they emphasize identifying the targets not mastered or possible reasons why. These are the six approaches we shall discuss:

1. **Profiling content areas strengths and weaknesses.** A deficit is defined as a student's low standing, relative to peers, in a broad learning outcome area in a subject. For example, a student may have less ability in subtraction and division than in addition and multiplication compared to peers.
2. **Prerequisite knowledge and skills deficits.** A deficit is defined as a student's failure to have learned concepts and skills necessary to profit from instruction in a course or a unit.
3. **Mastery of specific objectives.** A deficit is defined as a student's failure to master one or more end-of-instruction learning targets. Most so-called formative uses of state test data are in this category, for example, as results are used to group students as "below basic" on particular standards.
4. **Identifying students' errors in performance.** A deficit is defined as the type(s) of errors a student makes.
5. **Knowledge structure analysis.** A deficit is defined as a student's inappropriate or incorrect mental organization of concepts and their interrelationships.
6. **Component competencies of problem solving.** A deficit is defined as a student's inability to perform one or more of the components necessary to solve a word problem.

Each approach will be described and evaluated in terms of how well it meets the second purpose of diagnostic testing: identifying probable causes of a student's learning difficulty. Figure 7.1 illustrates each of the first four approaches with a specific example and serves as a tool for comparing the approaches. The last two approaches, illustrated later, are more in line with cognitively oriented instructional psychology.

Approach 1: Profiling Content Strengths and Weaknesses

In this approach, a school subject—say, elementary arithmetic or elementary reading—is subdivided into areas, each of which is treated as a separate trait or ability. *KeyMath3 Diagnostic Assessment* (Connolly, 2007), for example, divides mathematics into 3 areas (Basic Concepts, Operations, and Applications) and 10 subareas (numeration, algebra, geometry, measurement, data analysis and probability, mental computation and estimation, addition and subtraction, multiplication and division, foundations of problem solving, applied problem solving) and assesses a student in each area. Results are reported as a profile of strengths and weaknesses over the 10 subareas. As is typical of tests in this category, strengths and weaknesses are interpreted in norm-referenced ways: A student with a "weakness" is significantly below the norm. Percentile ranks (discussed in Chapter 16) are the

FIGURE 7.1 Examples of how different approaches to diagnostic assessment interpret the same student's performance.

Examples of items along with responses of a hypothetical student

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
17	15	43	337	654	43	63	562	667
−12	−13	−32	−226	−423	−25	−57	−453	−374
5	2	11	111	231	× 22	× 14	× 111	× 313

Total score for subtraction = 5/9 or 56%. Percentile rank = 18

Approach 1. Profile of strengths and weaknesses

The score on the subtraction subtest shown above is compared to the scores on other subtests such as addition, multiplication, division, etc. A profile of strengths and weaknesses in arithmetic is created for each student.

Example: The score of 5 correct has a percentile rank of 18 and is lower than other subtest scores.

Interpretation of the results: The student is weak in subtraction.

Approaches 2 and 3. Prerequisite hierarchy combined with mastery of specific objectives

The items above may be derived from a hierarchy of prerequisite arithmetic skills and the mastery of each skill in the hierarchy is assessed. (Skill statements are based on Ferguson [1970].)

Example:

	Hierarchy of Skills	*Score*
(4)	Subtract 3-digit numbers requiring borrowing from either tens' or hundreds' place. [Items (h) and (i)]	$^{0}/_{2}$ or 0 %
(3)	Subtract 2-digit numbers with borrowing from tens' place. [Items (f) and (g)]	$^{1}/_{2}$ or 50 %
(2)	Subtract two 2-digit and two 3-digit numbers when borrowing is not needed. [Items (c), (d), and (e)]	$^{3}/_{3}$ or 100 %
(1)	Subtract 2-digit numbers when numbers are less than 20. [Items (a) and (b)]	$^{2}/_{2}$ or 100 %

Interpretation of the results: The student has mastered the prerequisite Objectives 1 and 2, but has not mastered Objectives 3 and 4. Instruction should begin with Objective 3.

Approach 4. Identifying Errors

The subtraction item(s) that the student answered incorrectly are studied and the student's errors are identified.

Example: The student's responses to Items (f), (g), (h), and (i) are wrong. These are studied to identify the type(s) of errors the student made.

Interpretation of the results: The student is not renaming (regrouping) from tens' to units' place and from hundreds' to tens' place.

primary type of norm-referenced score used in this context.

Strengths This approach to diagnostic assessment is most useful to give you a general idea about students' performance in subareas of a subject matter. It fits with the intentions many states have of making large-scale test data "formative."

Weaknesses If the set of items that a test has to indicate performance on particular standard contains only a handful of items or tasks, the subtest scores probably will be unreliable. As a result, the students' strengths and weaknesses may be exaggerated or masked by chance errors of measurement.

Note that this approach does not tell you about attainment of particular learning goals in the absolute sense; rather, it gives relative strengths and weaknesses within the group. Diagnosis with such tests provides you with only general information about where their problems lie. It is much like saying, "The treasure lies to the north." The information is helpful, but it leaves you with a lot of work to do before the treasure can be found.

A good educational diagnostician will use the initial test results to formulate hypotheses concerning students' difficulties. You confirm or reject these hypotheses by following up and gathering additional information. Thus, although the initial profile of strengths and weaknesses may be unreliable, the final diagnosis will be much more reliable.

Approach 2: Identifying Prerequisite Deficits

This approach explores whether students have fallen behind because they have not learned the specific knowledge and skills necessary to profit from upcoming instruction. Among the approaches relying on identification of learning prerequisites are Gagné's **learning hierarchies** (Gagné, 1962, 1968; Gagné, Major, Garstens, & Paradise, 1962; Gagné & Paradise, 1961).

The first step in creating a hierarchy is to select one learning target the student must be able to perform. The next steps involve analyzing it to identify the prerequisites a student must learn in order to achieve the target. For each prerequisite identified, you repeat the same analysis, generating a hierarchy of prerequisite performances. This backward analytic procedure identifies critical prior learning, the lack of which could cause students problems in subsequent learning.

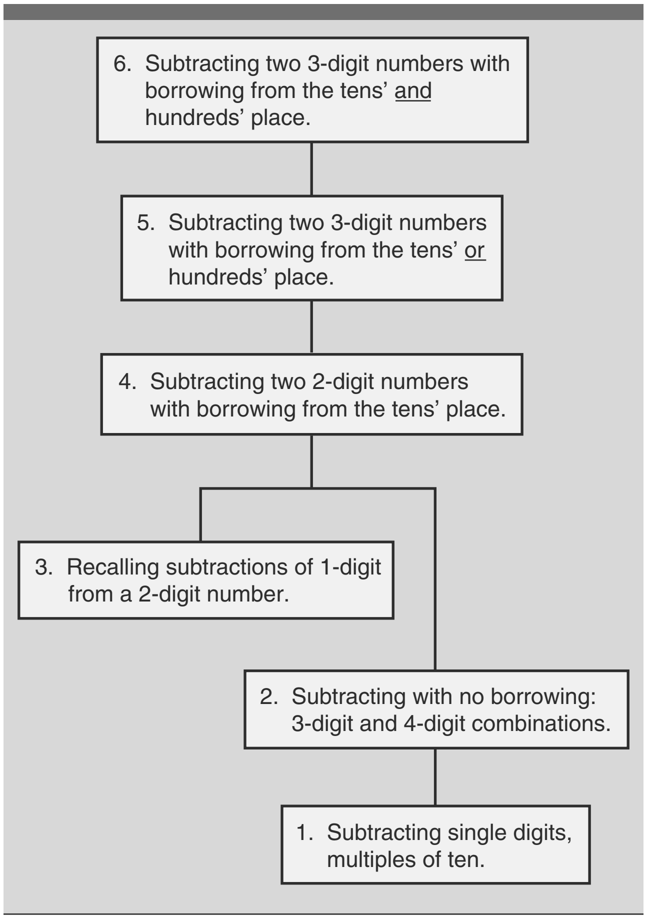
The difference between this approach and the previous one is that here you focus on whether each prerequisite was learned rather than on the pattern of profile strengths and weaknesses. Your interpretation of results is criterion-referenced rather than norm-referenced.

A contemporary rediscovery of the idea behind learning hierarchies can be found in a task analysis (building-block) approach to learning progressions (Popham, 2008). As we discuss below, not all approaches to learning progressions are diagnostic task analyses or building-block approaches, but this is one way to think about it. You identify prerequisites for a single unit of instruction or for a term. For each learning target, ask yourself, "What must students be able to do before they are ready to learn this learning target?" Focus on what needs to be learned immediately before. Once you identify that immediately prior (prerequisite) performance, ask the same question about it, and so on.

Figure 7.2 shows an example of a learning hierarchy for computational subtraction based on an analysis by Ferguson (1970). The final learning target, "subtracting two 3-digit numbers with borrowing from both the tens' and hundreds' place," is at the top. All the other performances are prerequisites to it. Notice that Performance 5 is prerequisite to 4, but that both 2 and 3 are prerequisites to 4. The 2 and 3 performances are not prerequisite to each other, however, so they are shown in parallel branches.

Once you have created the hierarchy, you assess each student with several items for each of the prerequisites identified. At a minimum, you should use four or five items per prerequisite in the hierarchy.

FIGURE 7.2 **Prerequisite hierarchy of a subtraction unit.**


6. Subtracting two 3-digit numbers with borrowing from the tens' and hundreds' place.
5. Subtracting two 3-digit numbers with borrowing from the tens' or hundreds' place.
4. Subtracting two 2-digit numbers with borrowing from the tens' place.
3. Recalling subtractions of 1-digit from a 2-digit number.
2. Subtracting with no borrowing: 3-digit and 4-digit combinations.
1. Subtracting single digits, multiples of ten.

Strengths This approach very specifically identifies skills that students need to learn before they are ready to be taught new learning targets. A hierarchy suggests the sequence for teaching the prerequisites. Assessments of prerequisite knowledge and skills are most helpful when you know very little about the students, especially when you expect large differences in their mastery of the prerequisites. Once you know each student's command of the prerequisites, you can tailor your teaching to meet his or her needs.

To be most effective, the prerequisites you identify should be specific to your curriculum and to your teaching approach. Different curricula and different teachers will approach the same subject differently. This means that the prerequisites for students you teach may be somewhat different than the prerequisites for students a colleague teaches. They should not be radically different, however, when both teachers are teaching comparable students the same material.

Weaknesses This approach is limited by the care and accuracy with which you analyze the learning requirements of your curriculum. If you do not identify the proper prerequisites, your assessment will lack validity. Further, in a continuous-progress curriculum, the distinction between prerequisites and "regular learning" is arbitrary, based more or less on instructional convenience.

The learning theory underlying this approach assumes that learning proceeds best by first teaching the prerequisites. This is a building-block approach, in which prerequisite performances build one on another to facilitate the learning of new targets. It is not clear that this building-block approach to learning is an appropriate teaching strategy for all subjects and for all students. Further, it does not provide information about *how* students understand or conceptualize their prerequisite knowledge.

Approach 3: Identifying Objectives Not Mastered

This approach centers the assessment on the important, specific targets students are expected to learn. Short tests assess each objective. The difference between this approach and the identifying prerequisite deficits approach is that here you assess only the objectives that are the outcomes of the unit or the course, not the prerequisite objectives.

The idea of teaching to specific learning objectives dates back at least to the seminal work of Waples and Tyler (Tyler, 1934; Waples & Tyler, 1930). Teachers in the 1930s and 1940s, however, generally did not write statements of behavior or develop diagnostic assessments based on them.

Earlier widespread popularity of behavioral objectives–based assessment and instruction was principally a result of the commercial availability of integrated sets of objectives and tests. Publishers moved toward such integrated materials in the late 1970s and early 1980s, after educators enthusiastically greeted the concept of criterion-referencing tests (Glaser, 1963; Nitko, 1989). Instructional methods were developed that integrated objectives, learning material, and diagnostic tests. The major prototypes were the mastery learning model (Bloom, 1968; Carroll, 1963), Individually Prescribed Instruction (Glaser, 1968; Lindvall & Bolvin, 1967), Program for Learning in Accordance with Needs (Flanagan, 1967, 1969), Individually Guided Instruction (Klausmeier, 1975), and Personalized System of Instruction (Keller, 1968; Keller & Sherman, 1974).

The diagnostic information you want to obtain from this approach is a list of learning targets (objectives) that students have and have not mastered. This means that for each teaching unit you must carefully identify and state the important learning targets. Follow these steps:

Step 1. Identify and write statements of the learning targets that are the main outcomes of the unit or the course.

Step 2. For each learning target, design four to eight test items.

Step 3. If possible, have another teacher review each item and rate how closely it matches the learning targets. Revise the items as necessary to obtain a closer match.

Step 4. Assemble the items into a single assessment instrument if the list of learning targets is relatively short (less than six); otherwise, depending on the students' educational development, you may need to divide the assessment into two or more instruments. For ease of scoring, keep all the items that assess the same objective together in one assessment.

Step 5. Set a "mastery" or passing score for each learning target. A frequently used **passing score** is 80% (or as near as you can come to this with the number of items you have for assessing a learning target). There is no educational justification for 80%, however. The important point is not the exact value of the passing score or passing percentage. Rather, it is the minimum level of knowledge a student needs to demonstrate with respect to each learning target to benefit from further instruction. This may vary from one learning target to the next. Use your own judgment, remembering that setting a standard too low or too high results in misclassifying students as masters or nonmasters.

Step 6. Administer the assessment to the students. After administering the assessment, separately score each learning target. Prepare a class list and chart in which you can record the students' scores on each learning target. This lets you identify students with similar deficits. Figure 7.3 shows an example of such a chart. The scores circled on the chart indicate a lack of mastery. Students should be given remedial instruction on these objectives.

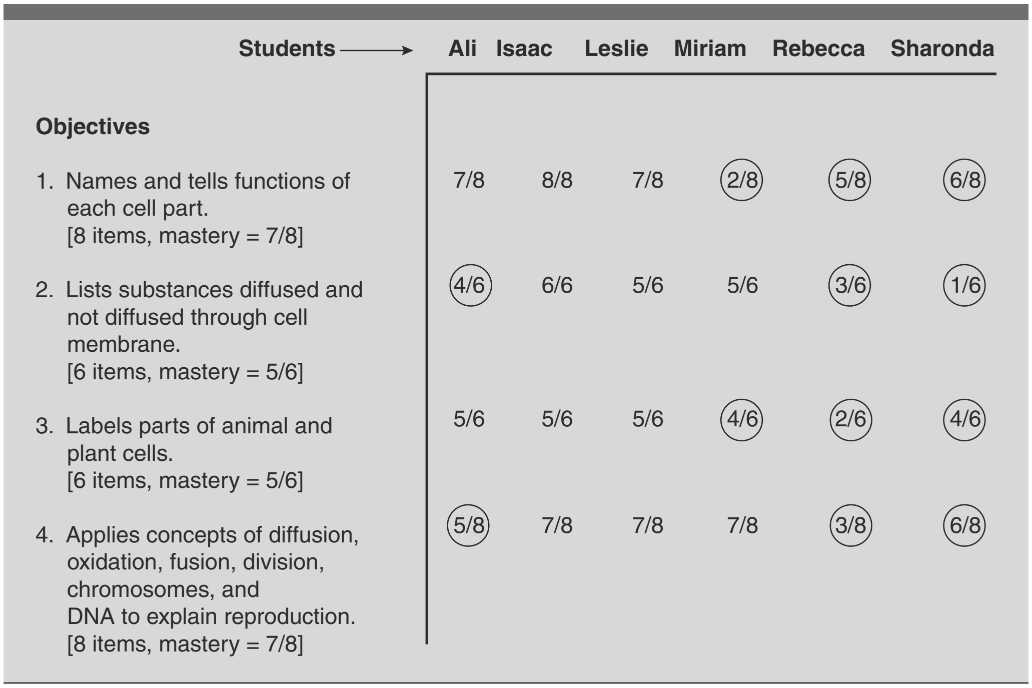
Strengths Diagnostic assessments based on specific objectives are appealing because they (a) focus on specific and limited learning targets to teach, (b) communicate learning targets in an easy-to-understand form, and (c) focus your attention on students' observed performance. These features make assessment easier, instructional decision making simpler, and public accountability clearer.

Weaknesses Objectives-based diagnostic assessments are generally plagued with measurement error, primarily because the assessments tend to have too few items per objective. If you use a diagnostic assessment to decide whether a student has "mastered" an objective, you should evaluate its quality using an index such as percentage agreement, rather than a traditional reliability coefficient. Percentage agreement is discussed in Chapter 4. A percentage agreement index estimates how likely students are to be classified in the same category when either the same assessment is readministered or an alternate form of the assessment is administered. (See Appendix J for examples of how to calculate this index.) Consistency of classification (i.e., of mastery or nonmastery) is the main focus, rather than consistency of students' exact scores.

The behavioral objectives approach to diagnostic assessment has other serious limitations. The information obtained reflects only one aspect of diagnosis: the observable behavior or performance of what is to be learned. This gives you little information about how to remediate the deficits you discover. You know only that a student has not mastered an objective. Like the other approaches we have discussed, behavioral objectives–based assessments are not fully diagnostic.

The behavioral objectives approach can also be criticized for implying an inappropriate theory of how knowledge and skills are acquired. A student's knowledge base is seen as a simple sum of previously learned specific behaviors. Further, critics point out that behavior-based tests fail to assess students' knowledge schemata, problem-solving disabilities, and abilities to think in new real-world contexts (Haertel & Calfee, 1983). Nor do they tap a student's internal representation (or schema) of knowledge, the relationships a student makes

FIGURE 7.3 Hypothetical example of diagnosis of specific objectives mastered and not mastered on a teacher-made, objectives-based diagnostic test. Circles mean nonmastery.



Students →	Ali	Isaac	Leslie	Miriam	Rebecca	Sharonda
Objectives						
1. Names and tells functions of each cell part. [8 items, mastery = 7/8]	7/8	8/8	7/8	2/8	5/8	6/8
2. Lists substances diffused and not diffused through cell membrane. [6 items, mastery = 5/6]	4/6	6/6	5/6	5/6	3/6	1/6
3. Labels parts of animal and plant cells. [6 items, mastery = 5/6]	5/6	5/6	5/6	4/6	2/6	4/6
4. Applies concepts of diffusion, oxidation, fusion, division, chromosomes, and DNA to explain reproduction. [8 items, mastery = 7/8]	5/8	7/8	7/8	7/8	3/8	6/8

between knowledge elements, the way students construct meaning from their learning experiences, and the knowledge-processing skills a student commands.

Finally, focusing on isolated and specific learning targets can make the curriculum seem fragmented. That is, the general themes and the learning goals that express integration of many specific knowledge and skill components are often neglected in favor of the isolated specific objectives.

Approach 4: Identifying Students' Errors

The goal of this approach is to identify student errors, rather than making a simple mastery-nonmastery decision about overall performance on a particular behavioral objective. Examples of errors are failure to regroup when "borrowing" in subtraction, improper pronunciation of vowels when reading, reversing *i* and *e* when spelling, and producing a sentence fragment when writing. Once you identify and classify a student's errors, you can attempt to provide instruction to remediate (eliminate) them.

Related to the error classification approach are methods that analyze complex performance into two or more component performances. If a student cannot perform the entire complex performance, diagnostic assessment identifies which component behaviors are lacking.

It is not easy to apply this approach because it takes considerable experience and skill to identify students' errors, and there may be more than one cause for an error. Consider the subtraction problems in Figure 7.1, for example. An inexperienced or unskilled teacher may not recognize the possible cause of the student's mistakes. Oftentimes, such teachers will say the student was "not careful" or "made careless errors." However, students' errors are rarely careless or random. Rather, *students' errors are often systematic*. Students may apply a rule or a procedure consistently in both appropriate and inappropriate situations. For instance, in Figure 7.1, the student appears to have consistently applied this rule: "Subtract the smaller digit from the larger digit." This rule works for problems (a) through (e), but does not work for Problems (f) through (i). *It is important, therefore, that you consider every error a student makes as having some systematic cause*. Try to identify what caused the error, or what rule the student is using, before you dismiss it as careless or random.

Interviewing students helps uncover many student errors. You can ask students to explain how they solved a problem, to explain why they responded the way they did, to tell you the rule for solving the problem, or to talk aloud as they go through the solutions to problems. These informal assessment procedures often reveal the types of errors a student is making. Chapter 11 discusses higher-order thinking and problem-solving assessment. Those assessment strategies are useful for discovering what types of problem-solving errors a student tends to make.

Strengths The chief advantage of the error classification approach over the behavioral objectives approach is that you discover not only *that* a learning target cannot be performed but also which aspects of the student's performance are flawed. This narrows your search for possible causes of poor performance. A skilled teacher can use this information to identify quickly one or more instructional procedures that have previously worked (remediated the error) with similar students.

Weaknesses Error classification procedures have serious drawbacks, however. There are several practical problems. Students make many different kinds of errors, and these are difficult to classify and to keep in mind while analyzing a particular student's performance. Frequently students demonstrate the same error for different reasons, so remedial instruction could be misdirected. Also, the amount of individual assessment and interpretation required seems prohibitive, given the amount of instructional time available.

More serious than practical problems of implementation, however, is the problem that if diagnosis only classifies errors, it still fails to identify the thinking processes a student has used to produce the errors. Just knowing the type of error (failing to borrow in subtraction) does not tell you the appropriate knowledge structures and cognitive processes a student needs to reach the desired outcome. Of course, cognitive analyses could be incorporated into error diagnostic procedures. We turn next to this possibility.

Approach 5: Identifying Student Knowledge Structures

A shortcoming of the diagnostic assessment approaches already mentioned is their strong ties to the **surface features** of subject-matter information and problem solving. Diagnosis should focus more on how students perceive the structure or organization of that content (i.e., the students' knowledge structures) and how they process information and knowledge to solve problems using that content knowledge. Frequently, students' everyday understandings of terms and phenomena are at odds with subject-matter experts' and textbooks' understandings.

One example of preinstructional assessment is our Chapter 6 discussion of a cold drink and a sweater. If you ask younger students what will happen to the temperature of a bottle of cold soft drink when it is wrapped up in a wool sweater, many will say that the sweater will warm the drink. In their schemata, "sweater" is something that Mother tells you to put on to keep warm. Thus, even though you may explain things clearly, they do not believe that a sweater has insulating properties that will keep a cold drink cold. You must relate the new concepts to their current thinking and schemata. You must help them reconstruct their knowledge structures. They need to understand how keeping their bodies warm and keeping the soft drink cold are linked by the concept of insulation. To believe it they need to understand the principles of insulation and how a sweater works as an insulator. You may need to conduct some experiments to support their new beliefs and understandings further.

Several methods are used to assess students' knowledge structures. These methods share the common perspective that as individuals become more proficient, their knowledge becomes more interconnected, more deeply organized, and more accessible. Probably the one most commonly used in classrooms is **concept mapping**. A concept map is a graphic way to represent how a student understands the relationships among the major concepts in the subject. An example of how a student might organize concepts related to a science unit on rocks is shown in Figure 7.4.

Notice that this concept map shows that the student has fairly well-organized knowledge of this unit's concepts. However, some important concept linkages are missing. For the most part, the student understands the concepts hierarchically (e.g., *granite* and *pumice* are included in the category called *igneous,* which is a type of *rock*). The student shows only one connection that is related to change or transformation of specific rocks or categories of rocks (*shale* changes to *slate*). The student can't fit into the map the concept *sediment* and so doesn't know that sediment can form *shale* or *limestone*. Other linkages are missing, too: Igneous rocks can weather and transform into sediment and sedimentary rocks; sedimentary rocks can form metamorphic rocks, which in turn can weather and return back to sedimentary rocks; and limestone can change into marble (Champagne & Klopfer, 1980).

FIGURE 7.4 Hypothetical example of a student's concept map of rocks.

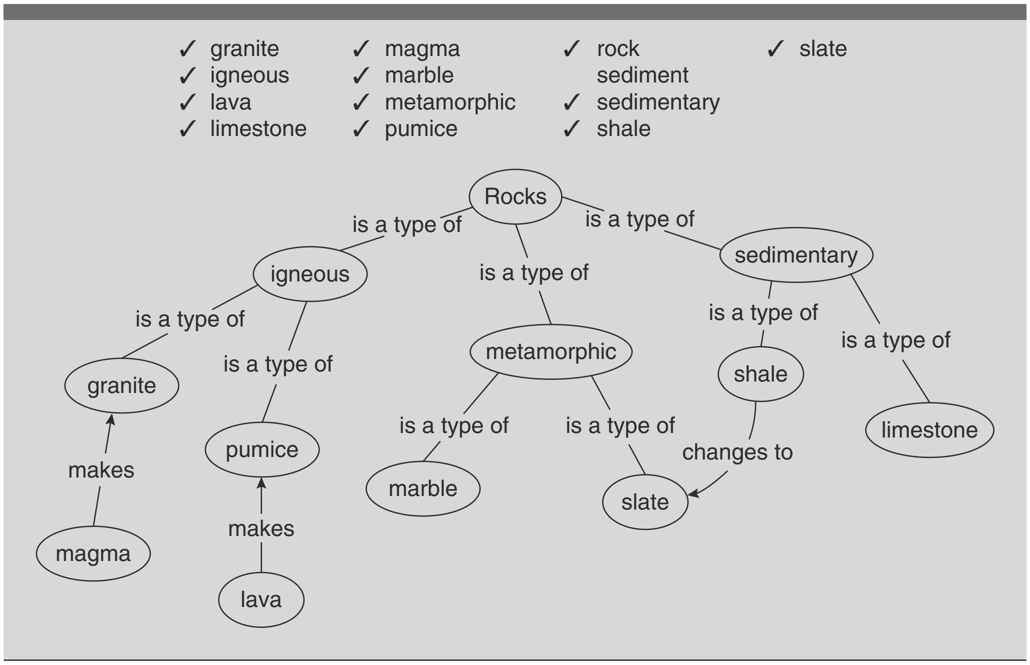


FIGURE 7.5 **Suggestions for conducting a student interview to create a concept map useful for assessment.**

Step	The focus of your interviewing and probing
1. Identify major concepts	Start with giving the student a few of the major concepts in the area you are probing. You could put these on cards. Ask the student to tell you about these concepts, what they mean to the student, and some of the other things about which they make the student think. Write every concept the student mentions in a list.
2. Create an arrangement of the concepts to match the student's thinking	Use a large sheet of paper. Review the list created in Step 1 with the student. Ask the student which of the concepts (including the ones you initially showed on the cards) is the major or most important one. Even if the student does not identify one as the major one, ask the student to pick one with which to start. Write this concept in the middle of the page. Ask the student to select another that is most closely related to the one on the page. Write this near the one already on the page. Continue asking for the ones nearest to the central one. Write these around the central concept. Continue with the remaining concepts, asking where they belong. These may be further from the central one and may be near some of the secondary concepts.
3. Establish how the student relates the concepts to one another	Begin with the central concept, work with the nearest ones to it, one at a time, and take each pair in turn. Ask the student whether the two are related and, if so, why or how they are related. Connect the related concepts with a line. Do not connect the concepts the student says are unrelated, even if you think they should be. Remember, you are trying to picture how the student is thinking. Assure the student frequently that there is no correct answer you are looking for but that you seek to help the student explain how he or she is thinking about these concepts. After connecting the related concepts with a line, write on the line the type of relationship the student tells you (e.g., "is an example of," "is a," "causes," "is part of," "it makes it go," etc.). If a student just says, "They are related," probe further to understand what the relationship is.
4. Give feedback to the student and rearrange the map	Show the student the map so far. Talk about the arrangement. Give feedback to the student, explaining what the map tells you about the student's thinking, and ask if this is correct. Rearrange the map so that it better represents the student's thinking and understanding of the concepts. Talk about each concept and its relationships. Add new concepts if the student mentions them and determine how they are linked to the mapped concepts. Redraw the map if necessary.
5. Elaborate the map to show new concepts, linkages, and examples	Further discuss the rearranged map with the student. Ask the student to tell you more: What else does the student know about these concepts, what are some examples, why are the concepts related, etc.? Incorporate this new information into the map and add branches and expansions as necessary to depict the student's thinking.
6. Explore cross-linkages and complex relationships	Go over the map drawn to this point with the student. Ask the student about the pairs of concepts previously unconnected and about the connections of new concepts mentioned in Step 5. Ask the student if he or she thinks three or four concepts should be connected together and why. Record these complex relationships.
7. Give feedback to the student and rearrange to make the final map	Show the map to the student and discuss with the student what the map tells you about the student's thinking. Ask the student if this is accurate and rearrange the map to make it more accurately describe the student's organization of the concepts. You may stop here if you have sufficient detail to understand the student's organization of the concepts. Otherwise, repeat Steps 5 and 6.

Suggestions for how to capture a student's concept map are given in Figure 7.5. For this task, the teacher shows a student the list of concepts at the top of Figure 7.4 and works with the student individually, following the procedure described in Figure 7.5, to create the concept map. As each concept is used in the map, it is crossed off the list.

As you can see from the example, using this approach to diagnosing requires individually assessing students, thoroughly knowing the subject so you can identify where a student has a missing link, and using considerable judgment when interpreting the resulting concept map. The validity of your judgments improves if you corroborate your assessment of a student's "missing links" with other evidence about how the student understands the concepts, such as problem-solving tasks and the student's essays and class responses. Also, keep in mind that there may be more than one correct way to relate the information; more than one schema may be correct.

Strengths This diagnostic approach focuses your attention on how a student thinks about the concepts and their interrelationships. It gives you some insight into how the student sees the concepts organized and, perhaps, how they might be related to other concepts and procedures a student has learned. These insights may help you explain why students are making errors, or why they are having difficulty solving problems.

Weaknesses Although assessment of knowledge structures and problem representation may offer you insight into a student's thinking, these procedures are experimental. We do not know the degree to which the results are valid, or whether different teachers would reach the same diagnosis for the same student. The way a student reacts to the teacher and the interviewing situation may drastically affect the results. You will need to be cautious, therefore, when you interpret the results. In large classrooms, these procedures present practical problems. Because you must assess one student at a time, you need to keep the rest of the class occupied. Although some of the procedures listed earlier are group oriented, it is not clear that they lead to the same results as individual interview methods.

Approach 6: Identifying Competencies for Solving Word Problems

This approach focuses on diagnosing whether students understand the components of word problems. Solving word problems comprises a significant number of learning targets in social studies, mathematics, and science. A word problem is a short verbal account of a more or less realistic situation that requires students to use the given information to answer a question. Consider the following word problem:

Example

A bus is carrying 38 passengers. It stops at a bus stop, where 23 passengers get off the bus and 11 other passengers get on. How many passengers are on the bus as it pulls away from this bus stop?

To solve this problem, a student must mentally process it using knowledge from long-term memory in several ways (Mayer, Larkin, & Kadane, 1984):

1. *Translation*—The student must understand each statement in the problem. This requires a student to use factual and **linguistic knowledge**.

Example

For example, in the preceding problem, a student must understand the concepts of *bus*, *carrying passengers*, *bus stop*, *get off the bus*, *get on the bus*, and *pulls away from this bus stop*. Linguistically, the student has to understand the meaning of the question, "How many passengers are on the bus as it pulls away from this bus stop?"

---

2. *Understanding*—The student must form a mental representation or model of the problem. In other words, the student must use **schematic knowledge** to recognize how the problem fits into a general framework to identify the type of problem it is. (See Chapter 11 for a discussion of schemata.)

Example

In the preceding problem, a student must recognize that this is an arithmetic problem involving only addition and subtraction.

---

3. *Planning*—The student must form a strategy or plan for solving the problem. The student must use **strategic knowledge**. (See Chapter 11 for a discussion of assessing solution strategies.)

Example

A student must recognize that to know how many passengers are on the bus as it leaves the bus stop, you must subtract from the 38 on the bus those 23 who got off at the stop and add to that remainder the 11 who got on at the stop. Arithmetically, the strategy is: $(38 - 23) + 11$.

(*Note:* Many problems may have more than one correct strategy.)

---

4. *Execution*—The student must use an appropriate algorithm (procedure) and carry out the calculations or steps properly. The student must use **algorithmic knowledge**.

Example

The student must be able to correctly calculate: (38 − 23) + 11 = 26 passengers on the bus as it leaves the bus stop.

The diagnosis in this approach is to identify students who are unable to solve word problems and whether their deficits lie in linguistic and factual knowledge, schematic knowledge, strategic knowledge, or algorithmic knowledge. A student may be unable to solve a problem because the student lacks one or more of these four types of knowledge. Your remedial instruction focuses on teaching students to use the type of knowledge in which they are deficient.

You apply this approach by identifying the critical types of linguistic, schematic, strategic, and algorithmic knowledge in each word problem. This means you must analyze each word problem and use the results of your analysis as a basis for asking diagnostic questions. Here is an example, similar to that used by Ismail (1994), of how you could phrase diagnostic items.

Example

Examples of diagnostic items for assessing students' knowledge of the component competencies of word problems

Focal Word Problem

The weight of an empty cookie tin is 3 ounces. When it is filled with cookies it weighs 1 pound. How many ounces do the cookies inside the tin weigh?

Linguistic Knowledge Diagnostic Items

1. What is a cookie tin?
2. What do you think the following question means: "How many ounces do the cookies inside the tin weigh?"

Schematic Knowledge Diagnostic Items

3. What are the arithmetic operations you need to solve this problem?

Strategic Knowledge Diagnostic Items

4. How many ounces are there in 1 pound?
5. What steps should you take to solve this problem? (Or, how would you go about solving this problem?)
6. Which of these is a correct way to solve this problem?
   A. (1 × 16) − 3
   B. (3 − 1) 16
   C. (1 × 3)
   D. 3 + 1
   E. 3 − 1

Algorithmic Knowledge Diagnostic Item

7. (1 × 16) − 3 =

The following suggestions will help you craft items for using with this type of diagnostic procedure.

1. *To assess linguistic knowledge,* focus your questions on the key terms and key phrases students must understand to translate the statement into a mental model of the problem. You may need to ask several questions and probe students' answers to discover their level of understanding of the words and phrases in a problem.
2. *To assess schematic knowledge,* ask students questions to see if they know which rules or principles they must use to solve the problem. For arithmetic problems, this may mean asking what operations should be used.
3. *To assess strategic knowledge,* focus on the students' ability to identify the proper sequence of steps or the proper processes needed to reach the answer. For arithmetic word problems, this means determining whether students know which numbers to use, which operations to use with those numbers, and the proper order of applying the operations. It may be easier to show students several sequences and ask which is the appropriate one for the given problem. All the numbers in the alternative solution strategies should relate to the word problem at hand.
4. *To assess algorithmic knowledge,* craft an item that presents the proper sequence and the proper numbers. The focus is on whether students can follow the algorithm without the context of the word problem. To avoid clueing the students as to the proper schema and strategy, present the algorithmic item after you have completed questioning for linguistic, schematic, and strategic knowledge.

Strengths This approach is most appropriate when you have word problems that are solved by applying a formula or a set of arithmetic operations in an algorithm. These include arithmetic word problems (such as money, time, rate, and cost), word problems in algebra, statistics word problems, social studies word problems involving mathematics, and science word problems. The framework you use to interpret the diagnosis (linguistic, schematic, strategic, and algorithmic knowledge) can be applied consistently across many categories of problems. The framework also suggests how you could remediate a student's deficits.

Weaknesses This approach requires many items per knowledge category to ensure sufficient reliability. Patterns you observe for one type of problem (e.g., money) may not emerge in other problem types (e.g., time). This makes diagnosis less valid if you try to generalize student deficits across problem types. Because many items are required, and because individual administration of items is the most appropriate assessment approach, the procedure is time-consuming. The validity of the approach also depends heavily on how well you are able to identify key phrases, appropriate schemata, and appropriate strategies for solving the problems. If multiple strategies for problem solving are appropriate, you must be careful to allow students to express these and not confuse the diagnosis by discounting them. Nevertheless, you can use this approach in a very informal way, perhaps asking questions orally, to get some insight into why students are having difficulty with word problems.

While the main purpose of diagnostic assessment is to support remediation of learning deficits, the main purpose of formative assessment is to support learning at all levels, improving on strengths as well as remediating weaknesses. And while the main emphasis in diagnostic assessment is to provide information for teacher planning, the main emphasis in formative assessment is to involve students in both generating and using assessment information. We turn to formative assessment in the next section.

FORMATIVE ASSESSMENT

Formative assessment loops assessment information back into the learning process itself. Effective formative assessment is based on learning goals and places a high value on appropriate teacher feedback and student self-assessment. Helping teachers develop effective formative assessment skills is the most cost-effective strategy for raising student achievement known today (Wiliam, 2007). This is because of its "double-barreled" nature, namely, that formative assessment addresses student cognitive and motivational needs at the same time.

This is what we mean when we say formative assessment is a loop: Students and teachers focus on a learning target, evaluate current student work against the target, act to move the work closer to the target, and repeat the process. This three-step process is an oversimplification, but it is a useful pattern to keep in mind for teaching and assessment (Sadler, 1983, 1989). In fact, if you had to only learn one thing about teaching, you might choose this cycle. From a student's point of view, the cycle is:

- What am I aiming for?
- How close am I now?
- What else do I have to do to get there?

The best formative assessment is student-centered, but it starts with the teacher's vision. First, you have to have the learning target clear in your own mind. This is not always as straightforward as it sounds. We once did an evaluation of a professional development program to teach middle school teachers how to assess reading. The middle school teachers had all been trained as English teachers, and their main areas of study had been literature and writing. Of course they knew what "reading" was, but they didn't understand that target well enough to help the students who reached middle school needing basic reading instruction. Without a detailed understanding of the target themselves, what they did with poor readers in their classroom was just "make them read" more, and their assessments indicated the students were—surprise—poor readers. The professional development program divided reading targets into five areas: oral fluency, comprehension, strategy use, higher-order thinking, and motivation. The idea was that the program could then offer assessment techniques for each of the five areas. According to teachers' evaluation interviews, the single best thing the program gave them was not the assessment techniques, but a clearer definition of what it meant to be a good reader. Students began to improve as their teachers became better able to show them what they needed to work on (fluency, comprehension, and so on).

Second, you have to communicate the target to students in ways that they understand. Typically, writing your objective on the board is not enough. Sometimes communicating your goals will involve showing students instead of telling them. For example, silent reading works better in elementary classrooms in which the teacher models silent reading, shares the books she reads, and talks about why she liked them than in classrooms in which the teacher uses silent reading time to catch up on paperwork.

Third, the students have to buy in. If you have been successful at communicating the target, you will have also helped students see why it is important for them to expend effort to reach it. This can be because of interest ("this topic is cool!") or academic, for example, when students are convinced to learn to write term papers in high school so they can do what is required in college. It can be because students want to be able to do something you or other adults can do, or something their older peers can do. Sometimes several of these motivations occur at the same time. For any given learning target, there will be a mixture of motivations in your class. For example, one student may be interested in a particular topic and another simply convinced that it is an important school target.

To properly communicate a target, you also need to share the criteria for good work. Otherwise, you and the students have no way to evaluate how close their work comes to being "good." You can do this by sharing criteria, for example, by giving students a copy of the scoring rubrics you will use to evaluate their final work. You can also show some examples of good work. Or, show some examples over a range of quality levels and let the students figure out what is "good" about the good work.

For some important assignments that you plan to use other years, ask some students if you can save a copy of their work to use in future classes. Most will be delighted. We know one teacher in Nebraska who saved "good example" copies of science notebooks each year to use with future students. She found that the quality level rose each year. Succeeding classes were able to grasp and meet, and then improve on, the standards of achievement shown in the notebooks.

Students can evaluate their own or peers' work against criteria you provide or criteria they deduce from examples, and provide feedback. Some research suggests that self-evaluation leads more directly to improvement than peer evaluation (Sadler & Good, 2006). You also should provide feedback, and we discuss particular ways to do that below.

Armed with appropriate feedback, students should have what they need to improve. For mastery learning targets, this is a more short-term and immediate process. (Practice today; find out what you need to work on; do better tomorrow.) For developmental learning targets like becoming a good writer, the process is longer. Students can take into account feedback on today's writing, but also on previous writing, when they write tomorrow.

Students should have the opportunity to evaluate their own learning. This is known as **student self-assessment**. Teach students effective self-assessment techniques; for many students, they don't come naturally. Offer opportunities for students to apply criteria to their own work in progress, discuss their work with peers, and reflect on their work after its completion.

We saw an especially striking contrast of the benefits of teaching self-assessment in two first-grade classrooms in a school district in Pennsylvania. All elementary students wrote reflection sheets to include in portfolios. One of the two first-grade teachers saw that her students had just filled in blanks on the reflection sheet, for example writing "Adding 5s" in the blank after "What did you learn?" because that was the title of the assignment sheet. She asked her students follow-up questions to stimulate further thinking (questions like "What did you learn to do when you add 5s?") and gradually got more reflective answers (such as "You get 5 or 0 in the ones place" or "I learned I [already] know it"). The other first-grade teacher just passed out the reflection sheets like worksheets, because the evaluation required it. Most of her students stayed in the "copying" phase. The difference between these classes was quite apparent to those of us who got to see both.

BENEFITS OF FORMATIVE ASSESSMENT

The effects of good formative assessment on achievement can be as much as 0.4 to 0.7 standard deviations—the equivalent of moving from the 50th percentile to the 65th or 75th percentile on a standardized test (Black & William, 1998). These effects exist at all levels—primary, intermediate, and secondary—and are especially noticeable among lower achievers. There are many reasons for these effects:

- Formative assessment helps teachers and students identify what students can do with help and what they can do independently.
- Participating in formative assessment is active learning, keeping students on task and focused on learning goals.
- Formative assessment, especially peer and self-evaluation, helps students with the social construction of knowledge.
- Formative assessment allows students to receive feedback on precisely the points they need in order to improve. It shows them what to do next to get better.

The latter reason is probably the most important.

Motivational benefits of formative assessment are a little more complicated. Different students respond differently to the various aspects of the formative assessment process.

Student self-assessment fosters both motivation and achievement. Students who can size up their work, figure out how close they are to their goal, and plan what they need to do to improve are, in fact, learning as they do that. Carrying out their plans for improvement not only makes their work better, it also helps them feel in control, and that is motivating. This process, called self-regulation, has been found to be a characteristic of successful, motivated learners.

Regulation of learning can be internal, as when students use self-assessment information to improve, or external, as when students use teacher feedback to improve. Either can support learning. Ideally, the internal and external work together.

The effects of feedback depend not only on the information itself but also on the characteristics of the people who send (teacher) and receive (student) the message. Whether students hear feedback as informational or controlling depends in part on them. One student may listen to a helpful, clear description of how to improve a paper with gratitude, while another may hear the same feedback as just another confirmation of how stupid he is. Covington (1992) wrote that while no two children come to school with equal academic abilities and backgrounds, there is no reason that they should not all have access to equally motivational feedback. He called this "motivational equity."

There is some evidence that good students use all information, including graded work, formatively (Brookhart, 2001). This is not the case for students who experience negative feelings after failure. These feelings get in the way of processing additional information about their learning. For them, the value of feedback is lost, overshadowed by the low grade. For unsuccessful and unmotivated students, you need to deal with negative feelings first, before providing other formative assessment information, in order to break the cycle of failure (Turner, Thorpe, & Meyer, 1998). For these students, formative feedback should begin with statements of accomplishment and suggest small, doable steps for improvement. And even such careful efforts don't always work, as the following true story from one of the authors shows.

Kasim was a poster child for the cycle of failure. Fifteen years old and in my seventh-grade English class, he never completed any assignment. He would write a line or two of an exercise or assignment, and then simply stop. Most of his teachers—including myself, I'm ashamed to admit—worked on getting him to "behave" first and learn second, so the class was not disrupted for the other students. Kasim lived in a foster home, had been abused as a child, and had the scars to prove it.

One day, in response to a brief writing assignment, Kasim brought me a three-page story, printed in tiny, cramped letters. It was an autobiographical story about how he had been separated from his sister, did not know where she was, and missed her terribly. It had a strong voice, expressive vocabulary, and readable (if not perfect) mechanics. I was excited. He could write! (I really hadn't been too sure about that.) More than that, he had wanted to write. Perhaps I got too excited, but for whatever reason, when I tried to encourage him and talk about his story, he appeared embarrassed to have written it and shut down. That was the first and last complete piece he did in a whole year.

Kasim would be a grown man now. When I think of him, I hope he's alive, I hope he's not in jail, and I hope he has found his sister. I'm not sure what could have broken his failure cycle or changed his negative attitude toward school. If I had it to do all over again, especially knowing what I know now about students like Kasim, I would have done things differently. I would have given him short assignments with more opportunities for peer and teacher feedback, and given him a whole lot more choice. Kasim's life was full of circumstances beyond his control, and with hindsight that included my class.

A BALANCED ASSESSMENT SYSTEM

Learning targets are the hub that connects

- assignments (which in embodying the learning targets serve to communicate them to students and to afford practice on them),
- teacher formative feedback and student self-assessment (which apprises the student of where he stands in relation to the learning targets and what he should do next),
- summative assessment (which evaluates the results of student efforts against the learning targets), and
- scoring criteria (which express the results of assessment in a symbol system designed to describe quality levels on the learning targets).

We have discussed the importance of assessments and scoring criteria matching learning targets, at both the content and cognitive levels, as a validity issue. The same principle of alignment holds for any classroom assignment. Students will interpret what you ask them to "do" (their assignments) as what you want them to learn. Thus all assignments, not just assessments, must embody the learning targets.

So for example, if the learning target is for students to write descriptive paragraphs, the assignments should include practice writing descriptive paragraphs. Formative feedback on these should be based on your criteria for "good" descriptive paragraphs. Students should have the opportunity to use the feedback. Finally, they write a descriptive paragraph that is graded according to those same criteria.

Formative assessments give you information about how long to "form" and when to "sum." When students' work gets close to the learning target, they are ready to demonstrate achievement on summative assessment. Students whose formative assessments show they don't need more practice, when classmates still do, can do enrichment work related to the learning target or use their time for some other work.

SOURCES OF FORMATIVE ASSESSMENT INFORMATION

Figure 7.6 provides some examples of formative assessment. We present this information with a major cautionary statement: Whether an assessment is "formative" depends on what you or the students do with the information obtained from it. Any assessment can be used formatively, and doing something you call a "formative assessment" without using the information to improve learning isn't formative at all. Perhaps the most powerful general formative assessment strategy is simply to get in the habit of asking students to give reasons for all their answers ("Why do you say that?"), whether correct or incorrect, and getting students in the habit of articulating what they know and where they think they're stuck.

Students and teachers should routinely share information about the quality of student work. Formative assessment activities typically allow an exchange of information by focusing on criteria for good work for a particular target. Conversations or observations can be just as important a source of information as finished work. For example, if you observe a student having trouble working at her desk, you know there is some problem. If you ask her where she is stuck, she may be able to give you enough information about her thinking that you can help her move along.

Many formative assessment activities involve putting student or teacher observations on paper where they are easy to see and then discuss. For example, some teachers routinely use reflections sheets. Or, some have students indicate by red light/green light or happy/sad faces on their work whether they are certain or uncertain about their understanding. It is easier to see and interpret a red light than to try to guess from students' expressions that they don't understand.

Formative assessment is not used for grading. Students need—and deserve—an opportunity to learn before they are graded on how well they have learned. Formative assessment is used before instruction, to find out where students are, and during instruction, to find out how they are progressing. The assessment is informational, not judgmental. Students are free to pay attention to figuring out how they are doing and what they need to work on without worrying about a grade.

Make formative assessment a part of your teaching. Plan your instruction in ways that provide opportunities for individual students to make

FIGURE 7.6 Examples of formative assessments.

Type of Learning Target	Formative Assessment(s)	Use of Results
Learning targets involving concepts	Students reflect on previous learning, attitude, and interest	Extending class discussion Selecting appropriate and interesting class activities Identifying and correcting misconceptions Building on previous knowledge (using no more review than is necessary)
Writing (e.g., descriptive, narrative, persuasive, or expository paragraph)	Peer editing Self-assessment and teacher conference	Revising Future writing Reflecting on why the revision is better than the first draft
Learning math tables, spelling words, and other “facts”	Students predict what study strategies (e.g., flash cards) will work best for them, and keep track of what works for them quiz by quiz Students record what they “know” and “don’t know,” gradually moving the “don’ts” into the “know” category as they progress	Students adjust own study strategies Students see exactly what they know and don’t, and have control over moving their own knowledge
Science or social studies content from textbooks	Students summarize reading in their own words, meet with a peer, and discuss how their summaries are alike/different Students make lists of vocabulary or concepts they feel they understand and those they find difficult	Extending class discussion Focus studying for unit test
Learning targets involving seatwork	Students have a “teacher alert” on their desks, turned to the happy face or the green light when they’re understanding and the sad face or red light when they need teacher help	Individual assistance in a “just in time” fashion, focused on the student-perceived source of difficulty
Learning targets involving classwork	Instead of questioning individual students, all students “vote” their answer so you can scan the class for understanding. Younger children can answer yes-no questions as a group by standing (“Stand up if you think that a soda wrapped in a sweater will get warm.”) Older students can use answer cards for multiple choice questions, or use electronic answer pads, or write one-minute responses on 3x5 cards	Adjust pacing of class instruction Adjust content of class instruction Extending class discussion Identifying and correcting misconceptions Building on previous knowledge (using no more review than is necessary) Understanding where all or most of the class is, not just a few students who have been called on
Learning targets involving projects or assignments graded with rubrics	Students look at examples of previous students’ work across a range of quality levels and discuss what makes the work of that quality Students “translate” the rubrics into their own words to make them “kid-friendly” evaluation tools Peer assessment of drafts or partial products Self-assessment and teacher conference	Improved understanding of the qualities of good work Revising and finishing the project or assignment Reflecting on the qualities of one’s own work for use in future work
Learning targets involving skills (e.g., reading aloud, using the library or computer, writing)	Students set and record a goal and work toward it Teacher suggests a goal, shares with students Observe students in the process of working (e.g., using a microscope) as well as the finished assignment	Students either realize goal (and set another) or can state how far they have come and what they still need to work on Adjust instruction at the individual or group level, as needed

formative decisions about their own learning. For example, provide self-assessment opportunities in your lesson plans. You also make formative decisions about the group's learning needs and provide group feedback. For example, you may return an assignment on which a large number of students demonstrated a misconception, and use the opportunity to reteach the material.

Teach students how to compare their performance with the learning target. Most students will not automatically reflect on their own work in the manner that you intend. For example, if you ask a student, "What did you learn?" without providing any guidance on what to do, many will copy the title of the assignment: "I learned two-digit subtraction" or "I learned how a bill becomes a law." Remember the first graders in our evaluation project.

Rubrics with clear performance level descriptions are helpful in this process. Even with good rubrics, however, students need instruction and practice in comparing their own work with the description in the rubric. Students can work together to compare their work to the learning targets. Teachers should provide a "safe" atmosphere for this, where criticism is seen as constructive and part of the learning process. That is an important lesson in itself.

There are some developmental differences in student use of self-evaluation. Younger children may focus on neatness and other surface characteristics of work when they first do self-evaluation (Higgins, Harris, & Kuehn, 1994). With instruction and practice, children can learn to focus on the learning target (Ross, Rollheiser, & Hogaboam-Gray, 2002).

Narrowing the gap between the student performance and a learning goal may not be a smooth process. Depending on the scope of the learning goal, you may need additional rounds of the formative assessment process for that goal. For example, students may write a series of essays in high school, each one benefiting from preceding teacher feedback and self-evaluations. No matter the scope of the accomplishment, students should be able to see their work getting closer to the goal, and should understand what specific feedback insights and learning strategies they used helped them close the gap. This is an empowering cycle.

LEARNING PROGRESSIONS

Experience or study will teach you the common misconceptions your students are likely to have along the way as they learn a particular concept. Knowing these, you will be able to more meaningfully evaluate performance levels and suggest next steps. **Learning progressions** are developmental sequences that describe typical progress in understanding or skill in a particular domain (Gong, 2008; Heritage, 2008; Hess, 2007).

Formative assessment works best when it is used in the context of a continuum, a vertical "picture" of what it means to learn or progress in a domain. This is very different from the approach to learning goals and objectives taken by most state standards and most curriculum materials (Heritage, 2008). A learning progression maps student progress in learning, not in accomplishing the teaching- and activity-based "goals" that sometimes form the learning targets for lessons or units of instruction.

Different researchers have taken slightly different approaches to learning progressions. Forster and Masters (2004) used progress maps that described typical growth in an area of learning, and which can be used by both classroom teachers and system evaluators to situate student learning on a continuum based in classroom instructional work. Popham (2008) has identified learning progressions with the task analysis approach that began in the era of behavioral objectives, which can be used by classroom teachers to plan lessons. Wilson and Draney (2004) described a system of progress variables—specific understandings and skills at a level of detail appropriate for classroom—that could be aggregated to the more general descriptions required for judgment of achievement of a state standard or curriculum goal.

An example of a learning progression in reading is found at the Website of the State of Victoria (Australia) Department of Education and Early Childhood Development (http://www.education.vic.gov.au/studentlearning/teachingresources/english/englishcontinuum/reading/default.htm). This is just one of the learning progressions of the department, which provides learning progressions in several disciplines. Others have made more specific learning progress variables, for example, in understanding forces and motions (Wilson & Draney, 2004). Learning progressions have been developed more in some curriculum areas than in others.

A major insight for formative assessment that learning progressions have given us is to focus on the "big picture" of learning, viewing students' work as points along a developmental continuum. This will help remove the blinders that can come with focusing too narrowly on students' successes or difficulties with a particular lesson activity, which is a real issue for classroom teachers whose instruction is, by definition, activity oriented.

Find a learning progression in your area, or construct a draft of one with colleagues, and see how mapping your students in this way helps you be more visionary in your selection of appropriate instruction and in giving appropriate feedback. Many teachers find, when trying to give feedback, that identifying what's wrong with a student's work and suggesting how to fix it comes much easier than identifying strengths in a student's work and suggesting how to build on them. Using a learning progression approach helps you see good work done in an assignment as more than an end of the road. We turn to feedback in the next section.

FEEDBACK

Formative assessment information for students comes mainly in the form of feedback. Good feedback is descriptive, specific, and contains information for improvement. The type of feedback you give should match the purpose you have for giving it. We illustrate different types of feedback here, so you will be better able to control the kind of feedback you give.

Feedback can vary according to the *kind of comparison* it makes.

- *Norm-referenced* feedback compares performance to other students. ("Your paragraph was the best in the class.")
- *Criterion-referenced* feedback compares performance to a standard and describes what students can or cannot do. ("You are particularly good at using a variety of descriptive adjectives.")
- *Self-referenced* feedback compares a student's performance to his own past performance, or sometimes to expected performance. ("This paragraph is better than the last one you wrote.")

The best formative feedback for practice work is criterion-referenced or self-referenced feedback. For students whose beliefs about their own capabilities are low, use self-referenced feedback to show them how they are improving.

Feedback can vary according to whether it describes *results* or *processes* underlying results.

- *Outcome feedback* is knowledge of results. ("You got a B on that paper.")
- *Cognitive feedback* describes the connections between aspects of the task and the student's achievement. ("It doesn't seem like you used the study guide very much.")

Cognitive feedback helps students know what to do to improve. Outcome feedback only supports improvement if students can internally generate the cognitive feedback (Butler & Winne, 1995). For example, a student may get back a paragraph on which the teacher marked three comma faults and conclude on his own, "I should study comma use." However, many students need the scaffolding provided when the teacher explicitly provides cognitive feedback. Suggest a short-term learning goal (what to aim for next), and suggest specific strategies the student can use to get there.

Feedback can vary according to its *functional significance* (Ryan, Connell, & Deci, 1985).

- *Descriptive feedback* gives information about the work. ("You developed your main character with lots of thoughtful details.")
- *Evaluative feedback* passes judgment on the work. (Giving an A or saying, "Good job!")

Descriptive feedback is more useful for formative assessment than evaluative feedback, because it has the potential to give students information they can use to improve. Check that the feedback you give students not only *is* descriptive, but that the descriptions are also statements of how the work relates to criteria you have shared with students.

Verbal feedback, whether oral or written, also varies in other ways that any verbal communication can vary.

- Feedback varies in *clarity*. Students have to clearly understand what your feedback means if it is to be useful to them.
- Feedback varies in *specificity*. General statements are usually less helpful for improvement than specific descriptions and suggestions.
- Feedback varies in *person*. First-person ("I" statements) feedback works for some formative feedback (e.g., "I don't understand what you mean here."). Third-person feedback can help you describe the work, not the student (e.g., "This

paragraph doesn't have supporting details" is better than "You didn't use supporting details"). Avoid second-person feedback. "You" did this or that comes out sounding like finger wagging.

- Feedback varies in *tone*. Keep the tone supportive. We know, for example, of one teacher who wrote, "You think like a chicken!" That's not helpful.

Not all students will hear feedback in the way you intend. For example, some students who have low self-efficacy or who are fearful may hear feedback you intended to be descriptive as evaluative. They may simply hear in your description a judgment that their work is "no good." Observe how students hear and respond to your feedback and what they do as a result.

Generally, more descriptive feedback is better for formative assessment. If the description only affirms what is good, however, it may not help students improve in the future. A good plan for written feedback on a student's paper is to describe a couple of positive aspects of the work and one aspect that needs improvement.

SYSTEMATIC RECORD KEEPING

Keep records of the important results of formative assessment, not for grading, but to keep yourself organized. For example, you should know what sort of feedback you have given, over time, to a student on a particular skill (e.g., writing). You can design your own class, individual, or group record-keeping sheets for specific purposes. You may wish to use a computer spreadsheet or database program.

Keeping records will help ensure that you are systematic and have an opportunity to observe all students on all the behaviors or skills you have decided are important. You will be able to see for which students you have observed target behaviors or skills, and make a point to observe the rest of them. Also, making notes will result in more complete and organized information than if you relied on your memory. Use patterns of observations to decide what each student needs, or what the group needs. If no natural opportunity to observe a skill presents itself, you may have to create one.

How many observations you want to see before you identify a pattern or draw conclusions will vary. For example, a kindergarten teacher might want to make sure she observes each child holding a pencil correctly at least five different times. Or a high school biology teacher might want to observe each student preparing a slide correctly at least twice.

CONCLUSION

This chapter has described diagnostic and formative assessment, both of which help inform teacher planning. Formative assessment should be directly helpful to students as well as teachers. It should help them decide about what and how to study, how to approach problems and other assignments, and how to develop learning strategies that work for them. In the next set of chapters (8 through 12), we discuss how to design and write test items, assessment tasks, and scoring schemes. As you read these chapters, remember that items or tasks themselves are not "formative" or "summative." The use of information—for further learning or for grading and other final decisions—determines that. You need high-quality information from well-designed assessment questions, items, and tasks for both uses.

EXERCISES

1. For a subject you teach or plan to teach, craft a diagnostic assessment procedure for Approaches 1, 2, 3, and 4. If there is time, try each approach with students who are experiencing learning difficulties. Revise your assessment procedure based on these student trials. Share the final versions of your assessment procedures with others in your course.
2. Each of these statements describes an instructional decision-making situation. Read each statement and decide the approach(es) to diagnostic assessment that may provide needed information.
   a. A teacher wonders whether Larissa missed several arithmetic story problems because she doesn't know her number facts.
   b. Trinh missed several addition computational problems involving mixed decimal fractions. His teacher wonders whether Trinh is counting the number of decimal places in each addend and using this count as the basis for placing the decimal point in the final answer.
   c. Lou missed several whole-number arithmetic problems involving carrying (regrouping). His

teacher wonders whether Lou has not remembered to add his "carries" to the sum of the digits in the next column.

d. Janet is a slow reader who frequently misses comprehension questions following a passage. Her teacher wonders whether Janet has reading reversals that cause her to misread some words in the passage.

3. For each of the following assessment activities, identify at least one formative use for the information the teacher will get from it. You may use Figure 7.6 to help you.

    a. Students set a "help" button on their desk to let the teacher know they're having trouble during math practice.

    b. Students get together in pairs to read and critique each others' reports on a planet.

    c. Students write new vocabulary words on flash cards and use a recipe box to file words into three categories: "know cold," "know most of the time," and "don't know."

    d. At the end of each social studies class, students write "one question I still have" on a 3 × 5 card and turn it in as their "ticket" out of class.

4. Identify the kind of feedback in each of the examples below.

    a. "I never want to see such sloppy work again!"

    b. "Use a capital A for Anne's name."

    c. "It was so wonderful to read your insightful description of Captain Ahab; I feel you really understand his motives."

    d. "All your spelling words were correct, so you get an extra 5 minutes at the computer."

Completion, Short-Answer, and True-False Items

KEY CONCEPTS

1. Align assessments to the content and performance requirements of your learning targets.
2. Short-answer items require a word, short phrase, number, or symbol response.
3. A true-false item consists of a statement or a proposition that a student must judge and mark as either true or false.
4. True-false items are very useful, because judging the truth of a proposition is important to thinking in any discipline. Most criticisms of true-false items are actually criticisms of poorly constructed true-false items.

IMPORTANT TERMS

partial credit

partial knowledge

proposition

random guessing

scoring key

short-answer varieties: association, completion, question

strip key

true-false varieties: correction, multiple true-false, right-wrong, true-false, yes-no, yes-no with explanation

verbal clues (specific determiners)

From Chapter 8 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

In this chapter we discuss how to craft simple forms of items suitable for paper-and-pencil quizzes and tests: short-answer and true-false items. When referring specifically to paper-and-pencil assessment tasks, we shall use the terms item and *test item*.

THREE FUNDAMENTAL PRINCIPLES FOR CRAFTING ASSESSMENTS

Any assessment should conform to three fundamental principles for crafting assessments:

1. Focus each assessment task entirely on important learning targets (content and performance).
2. Craft each assessment task to elicit from students only the knowledge and performance that are relevant to the learning targets you are assessing.
3. Ensure that each assessment task does not inhibit a student's ability to demonstrate attainment of the learning targets you are assessing by drawing on other, nonessential knowledge or skills.

The first principle is a strong one. Limit assessment tasks to those that focus on only educationally important learning targets. Assessing whether students have learned trivial performances or minor points of content is a waste of time.

To apply the second principle, you need a clear idea of what the learning target is. If a student has achieved the desired degree of learning, the student should complete the relevant assessment task correctly. If, on the other hand, a student has not achieved the desired degree of learning, the deficiency should also be apparent in the assessment results. Some poor assessment tasks elicit unwanted behaviors from students, such as bluffing, fear, wild guessing, craftiness, or testwise skills. Testwiseness is the ability to use assessment-taking strategies, clues from poorly written items, and experience in taking assessments to improve one's score beyond what one would otherwise attain from mastery of the subject matter itself (see Chapter 13 for more detail about testwiseness). These extra, unwanted behaviors may lead you to an inaccurate evaluation. Many of the suggestions in the next several chapters are specific ways to help you apply the second principle.

The third principle recognizes that imprecise wording in a question, for example, may make an item so ambiguous that a student who has the knowledge may answer it wrong. Similarly, simple matters such as inappropriate vocabulary, poorly worded directions, or poorly drawn diagrams may lead an otherwise knowledgeable student to respond incorrectly. Even the format or arrangement of an item on the page can inhibit some students from responding correctly. The third principle is amplified and applied to each item format discussed in this and the subsequent chapters.

Not all assessment experts would agree that there are only three basic principles, but most are likely to agree that these three are the important and fundamental principles for constructing classroom assessment tasks. These three encompass most of the specific suggestions that assessment experts have made over the years except, perhaps, those practical suggestions for efficient scoring.

SHORT-ANSWER ITEMS

Varieties of Short-Answer Formats

Short-answer items require a word, short phrase, number, or symbol response. There are three types of short-answer items: question, completion, and association (Wesman, 1971). The **question variety** asks a direct question and the students give short answers. (A question that requires the student to write paragraphs or longer responses is called an *essay item*. We will discuss essay items in Chapter 10.) Here are two examples:

Examples

Question Variety of Short-Answer Item

1. What is the capital city of Pennsylvania? (Harrisburg)
2. How many microns make up 1 millimeter? (1,000)

The **completion variety** presents a student with an incomplete sentence and requires the student to add one or more words to complete it. Here are two examples:

Examples

Completion Variety of Short-Answer Item

1. The capital city of Pennsylvania is (Harrisburg)
2. 4 + (6 ÷ 2) = (7) .

The **association variety** consists of a list of terms or a picture for which students have to recall numbers, labels, symbols, or other terms. This type of question is also called the *identification* variety. Here are some examples:

Examples

Association Variety of Short-Answer Item

On the blank next to the name of each chemical element, write the symbol used for it.

Element	*Symbol*
Barium	(Ba)
Calcium	(Ca)
Chlorine	(Cl)
Potassium	(K)
Zinc	(Zn)

Usefulness of Short-Answer Items

Abilities Assessed Short-answer items can assess students' performance of lower-order thinking skills such as recall and comprehension of information. The short-answer format also can be used to assess higher-level abilities such as the following:

1. Ability to make simple interpretations of data and applications of rules (e.g., counting the number of syllables in a word, demonstrating knowledge of place value in a number system, identifying the parts of an organism or apparatus in a picture, applying the definition of an isosceles triangle).
2. Ability to solve numerical problems in science and mathematics.
3. Ability to manipulate mathematical symbols and balance mathematical and chemical equations.

Figure 8.1 lists a large number of examples of short-answer items. As you will see in other chapters, multiple-choice and other objective items can also assess these abilities. The generic items from Figure 8.1 are not matched to specific learning targets. An item used directly from this table is unlikely to assess the learning target you have taught. Thus,

FIGURE 8.1 **Examples of short-answer items assessing different types of lower-order thinking skills.**

	Examples of generic questions*	**Examples of actual questions**
Knowledge of terminology	What is a _____? What does _____ mean? Define the meaning of _____?	What is a *geode?*
Knowledge of specific facts	Who did _____? When did _____? Why did _____ happen? Name the causes of _____.	What is the title of the person who heads the executive branch of government?
Knowledge of conventions	What are _____ usually called? Where are _____ usually found? What is the proper way to _____? Who usually _____?	What are magnetic poles usually named?
Knowledge of trends and sequences	In what order does _____ happen? Name the stages in _____. After _____, what happens next? Over the last _____ years, what has happened to _____? List the causes of the _____.	Write the life cycle stages of the moth in their correct order. 1st _____ 2nd _____ 3rd _____ 4th _____
Knowledge of classifications and categories	To what group do _____ belong? In what category would you classify _____? Which _____ does not belong with the others? List the advantages and disadvantages of _____.	Mars, Earth, Jupiter, and Venus are all _____

(*Continued*)

FIGURE 8.1 (*Continued*)

	Examples of generic questions*	**Examples of actual questions**
Knowledge of criteria	By what criteria would you judge _____? What standards should _____ meet? How do you know if _____ is of high quality?	What is the main criterion against which an organization such as Greenpeace would judge the voting record of a congressional representative?
Knowledge of methods, principles, techniques	How do you test for _____? When _____ increases, what happens to _____? What should you do to _____ to get the _____ effect?	Today the sun's rays are more oblique to Centerville than they were 4 months ago. How does Centerville's temperature today compare with its temperature 4 months ago?
Comprehension	Write _____ in your own words. Explain _____ in your own words. Draw a simple diagram to show _____.	What do these two lines from Shakespeare's Sonnet XV mean? "When I consider everything that grows, Holds in perfection but a little moment . . ."
Simple interpretations	Identify the _____ in the _____. How many _____ are shown below? Label _____. What is the _____ in _____.	In the blank, write the adjective in each phrase below. *Phrase* 1. A beautiful girl _____ 2. A mouse is a small rodent _____ 3. John found the muddy river _____
Solving numerical problems	(Problem statements or figures to calculate would be placed here.) Use the data above to find the _____.	Draw a graph to show John's activities between 2:00 p.m. and 2:45 p.m. ■ John left home at 2:00 p.m. ■ John ran from 2:00 p.m. to 2:15 p.m. ■ John walked from 2:15 p.m. to 2:30 p.m. ■ John sat from 2:30 p.m. to 2:45 p.m.
Manipulating symbols, equations	Balance these equations. Derive the formula for _____. Show that _____ equals _____. Factor the expressions below.	Balance this equation _____ $Cu + H_2SO_4 =$ _____ $CuSO_4 +$ _____ $H_2O + SO_2$

*The "blanks" in the generic items are for you to fill in. The generic items are simply suggestions to get you started. You generate your own items suitable for testing your students. Your items must match your learning targets to be valid.

you must review each item and match it to your learning targets to be sure it will function validly.

Strengths and Shortcomings The short-answer format is popular because it is relatively easy to construct and can be scored objectively. But short-answer items are not free of subjectivity in scoring. You cannot anticipate all possible responses students will make. Therefore, you often have to make subjective judgments as to the correctness of what the students wrote. Spelling errors, grammatical errors, and legibility tend to complicate the scoring process further. For example, to the question "What is the name of the author of *Alice in Wonderland?*" students may respond Carroll Lewis, Louis Carroll, Charles Dodgson, Lutwidge Dodgson, or Lewis Carroll Dodgson. Which, if any, should be considered correct? Although subjective judgment is proper, it does slow down the scoring process. It also tends to lower the reliability of the obtained scores.

An advantage of the short-answer format is that it lowers the probability of getting the answer correct by random guessing. A student who guesses randomly on a true-false item has a 50–50 chance of guessing correctly; on a four-option multiple-choice item, the student has one chance in four of randomly guessing the correct answer. For most short-answer items, however, the probability of randomly guessing the correct answer is zero. Short-answer items do not prevent students from attempting to guess the answer—they only lower the probability of the students guessing correctly.

In principle, guessing can be distinguished from using one's partial knowledge to help formulate an answer. Partial knowledge is not likely to result in the (exact) correct answer in short-answer items. Teachers, however, often give **partial credit** for responses judged to be partially correct. This is an appropriate practice and can result in more reliable scores *if* you use a **scoring key** that shows the kinds of answers eligible for partial credit. Using such a scoring key makes your assignment of partial credit more consistent from student to student, thereby improving reliability.

Creating Short-Answer Items

Short-answer items are easy to construct, but you must follow a few simple guidelines. The checklist summarizes these guidelines in the form of yes-no questions. Use this checklist to review items before you put them on your test. A no answer to any one question is sufficient reason for you to omit an item from tests until you correct the flaw. The guidelines are really applications of the three fundamental principles for crafting assessments. In the following paragraphs, we examine the checklist's guidelines in more detail.

✔ **CHECKLIST**

A Checklist for Reviewing the Quality of Short-Answer Items

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does the item assess an important aspect of the unit's instructional targets?
2. Does the item match your assessment plan in terms of performance, emphasis, and number of points?
3. If possible, is the item written in question format?
4. Is the item worded clearly so that the correct answer is a brief phrase, single word, or single number?
5. Is the blank or answer space toward the end of the sentence?
6. Is the item paraphrased rather than a sentence copied from learning materials?
7. If the item is in the completion format, is the omitted word an important word rather than a trivial word?
8. Are there only one or two blanks?
9. Is the blank or answer space in this item (a) the same length as the blank in other items, or (b) arranged in an appropriate column?
10. If appropriate, does the item (or the directions) tell the students the appropriate degree of detail, specificity, precision, or units you want the answer to have?
11. Does the item avoid grammatical (and other irrelevant) clues to the correct answer?

1–2. *The first two guidelines concern the importance of what is assessed and how the item matches the test blueprint.* Assess only important performance and content, and match tasks to your learning targets and the assessment plan. Even if you perform no other evaluation of your assessment, make it a habit to evaluate every test item using these two criteria.

3. *The question format is the preferred format for a short-answer item,* and is preferred over the completion format. Here's why: The completion format always implies a question. The student must read the incomplete sentence and mentally convert it to a question before answering. Therefore, the most straightforward thing to do is ask a direct question in the first place. Further, the meaning of the items is often clearer if you phrase them as questions instead of incomplete sentences. Consider how a completion item can be improved by converting it into a direct question:

Example

Poor: The author of *Alice in Wonderland* was __________.

Better: What is the pen name of the author of *Alice in Wonderland?* <u>(Lewis Carroll)</u>

Because the first version is not written in a question format, many correct answers are possible, including "a story writer," "a mathematician," "an Englishman," and "buried in 1898." The second version phrases the statement as a question, focusing the item on the specific knowledge sought.

As with all such rules, this one does have exceptions. Occasionally the question form of the item incorrectly suggests the need for a longer or more complex answer. In this case, the incomplete sentence serves better. Here is an example of how the question form of an item may imply a longer than necessary answer:

Example

Poor: Why are scoring guides recommended for use with essay tests?

Better: The main reason for using a scoring guide with an essay test is to increase the (objectivity) of the scoring.

Although the first version in the example implies that the teacher wants a paragraph or more, the teacher really had a very simple response in mind. This miscommunication is corrected by the second, revised version of the item. Most of the time, the question format produces better items. Your first impulse, therefore, should be to write questions, not incomplete sentences.

4. *Word the items specifically and clearly.* Usually, short-answer items require a single correct answer. You should word the question or incomplete sentence so this is clear to the student. Illustrations of how using the correct wording communicates that the teacher wants a single, specific answer include the following:

Example

Poor: Where is Pittsburgh, Pennsylvania, located? ____________________

Better: Pittsburgh, Pennsylvania, is located at the confluence of what two rivers? (Allegheny and Monongahela)

Better: What city is located at the confluence of the Allegheny and Monongahela rivers? (Pittsburgh, Pennsylvania)

Several answers to the first version are possible, depending on how specific you want the answer to be: "western Pennsylvania," "southwestern corner of Pennsylvania," "Ohio River," "Monongahela and Allegheny Rivers," and so on are all correct. If you want a specific answer, you must phrase the question in a focused and structured way. If you want to focus on the rivers, for example, the first rephrased version may be used. To focus on the city, use the second rephrased version.

Focusing the item is important because you want a certain answer. Some students who know the desired answer will not give it because they misinterpret the question. This is especially likely for students at the elementary levels who interpret questions literally. For example, in one classroom, fourth graders were given a bar graph to interpret. The teacher then asked the poorly phrased question in the example below:

Example

Poor: Was the population of Mexico greater in 1941 or 1951? ____________

One hapless student examined the graph and responded yes. We'll leave the revision of this item to you.

5. *Put the blank toward the end of the sentence.* This fifth guideline applies to completion items. If blanks are placed at the beginning or in the middle of the sentence, the student has to mentally rearrange the item as a question before responding to it. Even a knowledgeable student will have to read the item twice to answer it. The examples below show how to improve an item by putting the blank at the end:

Example

Poor: _________ is the name of the capital city of Illinois.

Better: The name of the capital city of Illinois is (Springfield).

Teachers of elementary-level arithmetic recognize that the ability to solve missing addend problems (e.g., "5 + ____ = 12" or "____ + 5 = 12") is quite difficult to learn. When blanks are not placed at the end of a sentence, the verbal item functions like these arithmetic problems. Unlike missing addend problems, however, putting blanks at the beginning of a sentence places an unintended barrier in the path of a youngster who has command of the relevant knowledge. Such barriers lower the validity of your assessments. Elementary students are sometimes observed stopping and puzzling at a blank without reading the entire item: They realize that they should write an answer there, but they lack the experience to read ahead and mentally rearrange the item as a question. If you rephrase the item as a direct question or place the blank at the end, these youngsters are able to display the knowledge they have acquired.

6. *Do not copy statements verbatim.* When you copy material, you encourage students' rote memorization rather than real comprehension and understanding. Further, textbook statements used as test items are usually quoted out of context. This may lead to item ambiguity or to more than one correct answer. One suggestion is to think first of the answer and then make up a question to which that answer is the only correct response.

7. *A completion item should omit important words and not trivial words.* Use the item to assess a student's knowledge of an important fact or concept. This means, for example, that you should not make the blanks the verbs in the statement. An exception, of course, would be a language usage item that focuses on the correct verb form.

8. *Limit blanks to one or two.* With more than one or two blanks, a completion item usually becomes unintelligible or ambiguous so that several unintended answers could be considered correct. Consider the following example:

Example

Poor: ________ and ________ are two methods of purifying ________.

Better: Two different methods of purifying water are (distillation) and (deionization).

9. *Keep all blanks the same length.* Testwise students sometimes use the length of the blank as a clue to the answer; avoid such unintended clues. When testing older students, you can save yourself considerable scoring time by using short blanks in the item and by placing spaces for students to record answers at the right or left margin of the paper or on a separate answer sheet. You can then lay a **strip key** with the correct answers along the edge of each student's paper and score papers quickly. Typing the items so that all blanks occur in a column accomplishes the same purpose. For example, instead of spreading the items across the page, you can arrange them as follows:

Example

Poor: Decisions for which rejection of some students is permitted are called ________ decisions.
Decisions for which every student must be assigned to one of several educational programs are called ________.

Better: Which type of educational decision permits rejection of some students?
____
Which type of educational decision requires that every student be assigned to one of several educational programs?
____

10. *Specify the precision you expect in the answer.* In a short-answer test involving dates or numerical answers, be sure to specify the numerical units you expect the students to use, or how precise or accurate you want the answers to be. This clarifies the task. It also saves time for students who strive for a degree of precision beyond your intentions. This example illustrates how to state the degree of precision expected in the answers:

Example

Poor: If each letter to be mailed weighs 1 1/8 oz., how much will 10 letters weigh? ______

Better: If each letter to be mailed weighs 1 1/8 oz., how much (to the nearest whole oz.) will 10 letters weigh? (11 oz.)

If there are more than one or two numerical items, you can describe the level of precision you expect in the general directions at the beginning of the set of questions, rather than adding words to each item.

11. *Avoid irrelevant clues.* A test item is designed to assess a specific learning target, but sometimes the wording provides an irrelevant clue. When this happens, a student may answer correctly without having achieved the learning target. The verb in a sentence, for example, may unintentionally clue the student that the answer you want is plural or singular. An indefinite article may be a clue that the answer you want begins with a vowel. The next example shows the same item with and without clues:

Example

Poor: A specialist in urban planning is called an (urbanist)

Better: A specialist in city planning is called a(n) (urbanist)

The poor version has two clues to the right answer: It uses *urban planning,* which clues *urbanist,* and it uses the indefinite article, *an,* which clues the student that the expected answer begins with a vowel sound. The better version corrects these flaws by substituting a synonym (*city planning*) and using *a(n)* for the indefinite article form.

TRUE-FALSE ITEMS

Varieties of True-False Items

A true-false item consists of a statement or a **proposition** that a student must judge and mark as either true or false. There are at least six **true-false varieties**: true-false, yes-no, right-wrong, correction, multiple true-false, and yes-no with explanation. The true-false variety presents a proposition that a student judges true or false. Here is an example:

Example

The sum of all the angles in any four-sided closed figure equals 360 degrees. T F

The **yes-no** variety asks a direct question, to which a student answers yes or no. This is an example:

Example

Is it possible for a presidential candidate to become president of the United States without obtaining a majority of the votes cast on election day?

Yes No

The **right-wrong** variety presents a computation, equation, or language sentence that the student judges as correct or incorrect (right or wrong). Here are two examples:

Example

Example assessing an arithmetic principle

$5 + 3 \times 2 = 16$ R W

Example assessing grammatical correctness

Did she know whom it was? C I

The **correction** variety requires a student to judge a proposition, as does the true-false variety, but the student is also required to correct any false statement to make it true. Here is an example along with the directions to the students:

Example

Read each statement below and decide if it is correct or incorrect. If it is incorrect, change the underlined word or phrase to make the statement correct.

The new student, who we met today, came from Greece. C I

The **multiple true-false** variety looks similar to a multiple-choice item. However, instead of selecting one option as correct, the student treats every option as a separate true-false statement. (More than one choice may be true.) Each choice is scored as a separate item. For example:

Example

Under the Bill of Rights, freedom of the press means that newspapers:

1. have the right to print anything they wish without restrictions. T F
2. can be stopped from printing criticisms of the government. T F
3. have the right to attend any meeting of the executive branch of the federal government. T F

The **yes-no with explanation** variety asks a direct question and requires the student to respond yes or no. In addition, the student must explain why his or her choice is correct. Here are some examples:

Example

Situation.*

A poll was taken of 500 city Democrats and 500 city Republicans. Each person was asked whether he or she agreed with the statement: "That government is best which governs least." These are the results:

Democratic male	Agree 12%	Disagree 35%
Democratic female	Agree 3%	Disagree 14%
Republican male	Agree 48%	Disagree 12%
Republican female	Agree 28%	Disagree 7%

1. I assert that this poll proves that most people want the government to do very little. Am I correct? Yes No
2. If you say no, explain why I am wrong: ______

*The situation portion of this item is adapted from Sanders, 1966, p. 117.

USEFULNESS OF TRUE-FALSE ITEMS

Advantages and Criticisms

Teachers often use true-false items because (a) certain aspects of the subject matter readily lend themselves to verbal propositions that can be judged true or false, (b) they are relatively easy to write, (c) they can be scored easily and objectively, and (d) they can cover a wide range of content within a relatively short period. But some educators have severely criticized true-false items—*especially poorly constructed true-false items*. Among the more frequent criticisms are that poorly constructed true-false items assess only specific, frequently trivial facts; are ambiguously worded; are answered correctly by random guessing; and encourage students to study and accept only oversimplified statements of truth and factual details. If you follow the suggestions in this chapter for improving true-false items, you can avoid these criticisms.

Assess More Than Simple Recall

Well-written true-false items can assess a student's ability to identify the correctness or appropriateness of a variety of meaningful propositions, including the following (Ebel, 1972):

1. *Generalizations* in a subject area
2. *Comparisons* among concepts
3. *Causal or conditional propositions*
4. *Relationships* between two events, concepts, facts, or principles
5. *Explanations* for why events or phenomena occurred
6. *Instances or examples* of a concept or principle
7. *Evidential statements*
8. *Predictions* about phenomena or events
9. *Steps* in a procedure or process
10. *Computations* (or other kinds of results obtained from applying a procedure)
11. *Evaluations* of events or phenomena

Examples of items of each of these types are shown in Figure 8.2. Some of the key phrases used to construct items in each category appear as well. You may want to refer to Figure 8.2 from time to time to glean suggestions for writing true-false items. The final item should assess your intended thinking skill and learning target. Using a key phrase from the figure does not guarantee that the item you craft will assess the thinking skill shown; check to make sure that it does.

Validity of the True-False Item Format

Ebel, perhaps more than any other measurement specialist, defended the use of well-written true-false items for classroom assessment. He offered the following argument for the validity of this format:

1. The essence of educational achievement is the command of useful verbal knowledge.
2. All verbal knowledge can be expressed in propositions.
3. A proposition is any sentence that can be said to be true or false.
4. The extent of a student's command of a particular area of knowledge is indicated by his success in judging the truth or falsity of propositions related to it. (Ebel, 1972, pp. 111–112)

Requiring students to identify the truth or falsity of propositions is not the only means of ascertaining their command of knowledge. Other ways of assessing command of knowledge are discussed further in the remaining chapters of this book.

Guessing on True-False Items

A common criticism of true-false tests is that they are subject to error because students can answer them with random guesses. It is well known that for a single true-false item, there is a 50–50 chance of answering the item correctly if true or false is selected at random. This means that persons guessing randomly can expect to get *on the average* one half of the true-false items correct. Several points, however, blunt this criticism (Ebel, 1972):

1. Blind (completely random) guessing is quite unlike informed guessing (guessing based on partial knowledge).
2. Well-motivated students tend to guess blindly on only a small percentage of the questions on a test.
3. It is very difficult to obtain a good score on a test by blind guessing alone.
4. If a given true-false test has a high reliability coefficient, that would be evidence that scores on that test are not seriously affected by blind guessing.

FIGURE 8.2 **Types of statements that could form the basis for your true-false items.**

Type of statement	Examples of introductory words or phrases	Examples of true-false items
Generalization	All . . . Most . . . Many . . .	All adverbs modify verbs. (F)
Comparative	The difference between . . . is . . . Both . . . and . . . require . . .	Both dependent and independent clauses contain subjects and verbs. (T)
Conditional	If . . . (then) . . . When . . .	When there is no coordinating conjunction between two independent clauses, they should be separated by a colon. (F)
Relational	The larger . . . The higher . . . The lower . . . Making . . . us likely to . . . Increasing . . . tends to . . . How much . . . depends on . . .	The amount of technical vocabulary you should include in an essay depends on your intended audience. (T)
Explanatory	The main reason for . . . The purpose of . . . One of the actors that adversely affect . . . Since . . . Although . . .	One of the factors affecting changes in rules governing English grammar and style is changes in how people use the language. (T)
Exemplary	An example of . . . One instance of . . .	The movie title *The Man Who Came to Dinner* contains a nonrestrictive clause. (F)
Evidential	Studies of . . . reveal . . .	Studies of contemporary literature show that some authors deliberately violate style and usage rules to create literary effects. (T)
Predictive	One could expect . . . Increasing . . . would result in . . .	Increasing the number of clauses in sentences usually increases the reading difficulty of a passage. (T)
Procedural	To find . . . one must . . . In order to . . . one must . . . One method of . . . is to . . . One essential step . . . is to . . . Use . . . of . . . The first step toward . . .	The first step toward composing a good essay is to write a rough draft. (F)
Computational	(Item includes numerical data and requires computation or estimation.)	There are two adjectives in the sentence "The quick brown fox jumped over the lazy dog." (F)
Evaluative	A good . . . It is better to . . . than . . . The best . . . is . . . The maximum . . . is . . . The easiest method of . . . is to . . . It is easy to demonstrate that . . . It is difficult to . . . It is possible to . . . It is reasonable to . . . It is necessary to . . . in order to . . . The major drawback to . . . is . . .	It is generally better to express complex ideas as two or more shorter sentences rather than one longer sentence. (T)

Note: To be valid, the items must match specific learning targets.

Source: Adapted from *Essentials of Educational Measurement* (pp. 183–185), by R. L. Ebel, 1972, Upper Saddle River, NJ: Prentice Hall. Adapted by permission of the copyright holder.

Random guessing, of the type that is assumed by the "50–50 chance" statement, is by definition random responding. Random guessing is sometimes called *blind guessing*. But most everyone's experience is that students rarely respond this way to test questions. Rather, students tend to use whatever **partial knowledge** they have about the subject of the questions and/or about the context in which the questions are embedded to make an informed guess. Such informed guessers have a higher than 50–50 chance of success on true-false items (but how much higher, we are unable to say). This means that scores from true-false items (as with other item types) are measures of partial knowledge when informed guessing occurs. (Of course, persons who actually know the answer have a 100% chance of answering correctly!)

Although a student who is responding randomly on a single true-false item will have a 50–50 chance of being correct, the laws of chance indicate that the probability of getting a good score by random guessing on a test made up of many true-false items is quite small, especially for longer tests. This is illustrated in Figure 8.3. Chances are only 2 in 100, for example, that a student who has guessed randomly on all the items on a 15-item test will get 80% or more of the items correct. If the test has 20 true-false questions and a student guesses randomly on all items, that student has only 2 chances in 1,000 of getting 80% or more items correct. Chances of a perfect (100% correct) paper are even smaller.

Suggestions for Getting Started Properly

To write good true-false items, you must be able to identify propositions that (a) represent important ideas, (b) can be defended by competent critics as true or false, and (c) are not obviously correct to persons with general knowledge or good common sense who have not studied the subject (Ebel, 1972). These propositions are then used as starting points to derive true-false items.

In this regard, Ebel (1972) suggested that you think of a segment of knowledge as being represented by a paragraph; the propositions are the main ideas of that paragraph. You can then use these main ideas as starting points for writing true-false items. Figure 8.2 offers suggestions on how to get started in phrasing true-false items from these propositions.

Frisbie and Becker (1990) offer these additional suggestions for getting started:

1. *Create pairs of items, one true and one false, related to the same idea, even though you will use only one.* Creating pairs of items helps you check on a statement's ambiguity and whether you need to include qualifications in the wording. Frisbie and Becker suggest that your false item is not worth using if you can only write a true version of it by inserting the word *not*.

2. *If your statement asks students to make evaluative judgments ("The best . . . is . . . ," "The most important . . . is . . . ," etc.), try to rephrase it as a comparative statement ("Compared to . . . , A is better than . . .").* The comparative statement allows you to put into the item itself the comparisons you want students to make. Usually, when you write "What is the best way to . . . ," for example, you raise in the mind of the student the question, "compared to what?" Thus, if you include your intended comparison in the statement itself, this clears up the ambiguity.

3. *Write false statements that reflect the actual misconceptions held by students who have not achieved the learning targets.* To do this, you have to know your students well and try to think about a proposition

FIGURE 8.3 Chances of a student obtaining various "good scores" by using only random guessing for all items on true-false tests of various lengths.

Number of T-F items on the test	Chances of getting the following percentage of T-F questions right:		
	60% or better	80% or better	100%
5	50 in 100	19 in 100	3 in 100
10	38 in 100	6 in 100	1 in 1,000
15	30 in 100	2 in 100	3 in 100,000
20	25 in 100	6 in 1,000	1 in 1,000,000
25	21 in 100	2 in 1,000	3 in 100,000,000

Note: Computations are based on binomial probability theory.

the way a misinformed or poorly prepared student thinks about it. As you teach, you may notice these misconceptions. Take notes so you can recall them as you write items.

4. *You may wish to convert a multiple-choice item into two or more true-false items.* The foils (or incorrect options) of a multiple-choice item may be used as a basis for writing false statements.

Suggestions for Improving True-False Items

Review and revise the first drafts of all your assessment tasks. Editing assessment tasks is an important step in the assessment development process. The checklist summarizes principles for improving the quality of true-false items. These are written as questions you can ask when you review your item drafts. You should always use the checklist to review true-false items that come with your textbook and curriculum materials, because these true-false items are notorious for their poor quality. The principles implied by the checklist are explained and illustrated in the following section.

✔ **CHECKLIST**

A Checklist for Judging the Quality of True-False Items

Revise every item for which you answered no to one or more questions.

1. Does the item assess an important aspect of the unit's instructional targets?
2. Does the item match your assessment plan in terms of performance, emphasis, and number of points?
3. Does the item assess important ideas, knowledge, or understanding (rather than trivia, general knowledge, or common sense)?
4. Is the statement either definitely true or definitely false without adding further qualifications or conditions?
5. Is the statement paraphrased rather than copied verbatim from learning materials?
6. Are the word lengths of true statements about the same as those of false statements?
7. Did you avoid presenting items in a repetitive or easily learned pattern (e.g., TTFFTT . . . , TFTFTF . . .)?
8. Is the item free of verbal clues that give away the answer?
9. If the statement represents an opinion, have you stated the source of the opinion?
10. If the statement does not assess knowledge of the relationship between two ideas, does it focus on only one important idea?

1–2. *The first items on the checklist cover the importance of what is assessed and its match to the test blueprint.* As always, the first two criteria that your assessment tasks should meet are importance and match to your assessment plan. Eliminate every item failing to meet these two criteria.

3. *Assess important ideas, rather than trivia, general knowledge, or common sense.* Although this guideline applies to all assessment tasks, you need to be especially sensitive to this point when writing true-false items. It is easy to write items that assess trivial knowledge. Here are some examples of how to improve items so they focus on more important ideas:

Example

Poor:	George Washington had wooden teeth.	T	F
Better:	George Washington actively participated in the Constitutional Convention.	T	F

The poor item focuses on trivia rather than important information about Washington's role in the early days of the nation. The revised version at least asks a more significant fact about him.

4. *Make sure the item is either definitely true or definitely false.* A proposition should not be so general that a knowledgeable student can find exceptions that change the intended truth or falsity of the statement. Make sure the item is phrased in a way that makes it unambiguous to the *knowledgeable* student. (Items should, of course, appear ambiguous to the unprepared or unknowledgeable student.) A few suggestions for reducing item ambiguity include the following:

a. *Use short statements whenever possible.* This makes it easier to identify the idea you want the student to judge true or false. Complex, cumbersome statements make identifying the essential element in the item difficult even for knowledgeable students. If the information you want to describe in the statement is complex, use different sentences to separate the description from the statement students must judge true or false. Frequently, you can shorten a long, complex statement that contains extraneous material by simply editing it.

b. *Use exact language.* Frequently, quantitative terms can clarify an otherwise ambiguous statement. For example, instead of saying "approximately $5.00" or "approximately one half of . . . ," say "between $4.00 and $6.00" or "between 45% and 55%."

c. *Use positive statements and avoid* double negatives, which many students find especially confusing. Here are some examples of improving items by avoiding negatives and double negatives:

Example

Poor:	The Monongahela River does not flow northward.	T	F
Better:	The Monongahela River flows southward.	T	F

If you must use a negative function word, be sure to underline it or use all capital letters so it is NOT overlooked. Do not take a textbook sentence and make it false by adding a "not" to it. The practice of taking a textbook sentence and making it false by inserting negative function words (e.g., *not, neither, nor*) makes the item you write tricky for students.

5. *Avoid copying sentences verbatim.* Students often find sentences copied from a text uninterpretable because they have been taken out of context. In addition, such statements are likely to communicate to students that the text's exact phrasing is important, rather than their own comprehension. This encourages students to engage in rote learning of textbook sentences. Recall from Figure 3.2 that one factor in the validity of your assessment is that it does not have such negative consequences.

Copying items from a text is more likely when a teacher is testing for knowledge of verbal concepts (including definitions) and statements of principles (rules). But testing for comprehension demands paraphrasing at the minimum, and enhancing a student's comprehension of concepts and principles seems to be a more important educational goal than encouraging a student to memorize textbook statements word-for-word.

6. *True and false statements should have approximately the same number of words.* Teachers tend to make true statements more qualified and wordy than false statements. Testwise students can pick up on this irrelevant clue and get the item right without achieving the learning target. Keep a watchful editorial eye and rewrite inappropriate statements.

7. *Don't present items in a repetitive or easily learned pattern* (e.g., TFTF . . . , TTFFTT . . . , TFFTFF . . .). Some teachers develop such patterns because they are easy to remember and thus make scoring easier. But if it's easy for a teacher to remember, it will also be easy for testwise students to learn. Assessment results will then be invalid. You should also avoid a consistent practice of having many more true answers than false, or many more false answers than true. If students notice, for example, that you seldom use a false statement, they will (rightly) avoid choosing false when they are uncertain of the answer. Upper-grade students discover these patterns quickly when a teacher uses lots of true-false items.

Not all educational assessment specialists agree on the proportion of true-to-false answers to include (Frisbie & Becker, 1990). Some specialists (Ebel & Frisbie, 1991; Popham, 1991) recommend having more false items than true ones, because false items have been shown to discriminate (distinguish between more and less knowledgeable students; see Chapter 13) better than true items (Barker & Ebel, 1981). *Discriminate* in this context means that false items tend to differentiate the most knowledgeable students from the least knowledgeable better than true items. Increased item discrimination improves the reliability of the total test scores.

8. *Do not use* **verbal clues** *(specific determiners) that give away the answer.* A **specific determiner** is a word or phrase in a true-false or multiple-choice item that "overqualifies" a given statement and gives the student an unintended clue to the correct answer. Words such as *always, never,* and *every* tend to make propositions false. Words such as *often, usually,* and *frequently* tend to make propositions true. Testwise students will use these clues to respond correctly even though they do not have command of the requisite knowledge. Here is an example of a poor item using a specific determiner:

Example

Poor:	In a ground war, the army with more sophisticated weaponry always defeats its opponents.	T	F

9. *Attribute the opinion in a statement to an appropriate source.* If your true-false item expresses an opinion, value, or attitude, attribute the statement to an appropriate source. You can use an introductory clause, such as "According to the text . . ." or "In the opinion of most specialists in this area . . ." or "In Jones's view. . . ." This referencing reduces ambiguity in two ways: (a) it makes clear that the statement is not to be judged in general, but rather in terms of the specific source; and (b) it makes clear that you are not asking for the student's personal opinion.

10. *Focus on one idea.* Have only one idea per item, unless the item is intended to assess knowledge of the relationship between two ideas. The following example shows how an item can be improved by focusing it on only one idea:

Example

Poor: The Monongahela River flows north to join the Allegheny River at Columbus, where they form the Ohio River. T F

Better: The Monongahela River and the Allegheny River join to form the Ohio River. T F

In the poor item, a student may respond with the correct answer, F, for an inappropriate reason: The student may think (erroneously) that the Monongahela River does not flow north, may be unaware that the confluence of the rivers is at Pittsburgh, or may lack any knowledge about the three rivers. Thus, the student would get the item right without having the knowledge that getting the right answers implies. A separate statement for each idea may be necessary to identify precisely what the student knows.

Creating Multiple True-False Items

A multiple true-false item looks like a multiple-choice item in that it has a stem followed by several alternatives. Unlike when responding to a multiple-choice item, however, the student does not select the single correct or best answer; she responds true or false to every alternative. In turn, each alternative is scored correct or incorrect. Because of this, the item may be constructed to have several correct (true) alternatives instead of only one. The examples that follow illustrate this.

Examples

**Under current collective bargaining laws, workers have the right to bargain for

1. how much workers of each skill level should be paid. T F
2. how much managers should be paid. T F
3. what new products the company should produce. T F
4. which workers should be laid off first. T F

**The following statements are arguments used by some people before the Civil War to justify slavery. Decide whether each statement is an argument based on democratic ideas.

1. Slavery is right because it existed through most of history. Y N
2. Slavery is right because the men who wrote the U.S. Constitution accepted it. Y N
3. Slavery is right because the great Greek, Aristotle, supported it. Y N

Source: Adapted from Sanders, 1966, pp. 53 & 132.

Format Notice three things about the format of the preceding examples. First, unlike multiple-choice items, the *options are numbered* consecutively, and asterisks set off the different clusters' stems. Second, you do not need to have a balance of true or false correct answers within one cluster. Some clusters, like the second one, may not have any true or any yes answers. Third, all of the statements within a cluster must relate to the same stem or question. Each statement within a cluster is treated as a separate true-false item. Thus, the example contains seven items, not two items.

Advantages This item format has the following advantages (Ebel & Frisbie, 1991; Frisbie, 1992): (a) Students can make two or three multiple true-false responses in the same time it takes them to answer one multiple-choice item; (b) a multiple true-false test created from multiple-choice items has a higher reliability than the original multiple-choice test; (c) multiple true-false items can assess the same abilities as straight multiple-choice items that are crafted to assess parallel content; (d) students believe that multiple true-false items do a better job of assessing their knowledge than straight multiple-choice items; (e) students perceive multiple true-false items to be slightly harder than straight multiple-choice items; and (f) multiple true-false items may be easier to write than multiple-choice items because you are not limited to creating only one correct answer.

Limitations The multiple true-false item format shares many of the same limitations as multiple-choice items. These limitations are discussed in Chapter 9. Some research shows that standard multiple-choice items may be more appropriate than multiple true-false items for assessing higher-order thinking skills and when criterion-related validity evidence is important (Downing, Baranowski, Grosso, & Norcini, 1995).

CONCLUSION

In this chapter, we discussed how to write effective short-answer and true-false test items. We summarized these item-writing principles in checklists. Some of these principles—assess an important aspect of the unit's instructional targets; match your assessment plan in terms of performance, emphasis, and number of points; and use clear, concise written expression—are principles for writing all types of test items, and some of the principles are unique to the item genre. We continue with two more item types, multiple-choice and matching, in Chapter 9.

EXERCISES

1. Write short-answer or completion items in your teaching area(s) that assess each of the lower-order thinking skills listed in Figure 8.1.
2. Each of the following completion items contains one or more flaws. For each item, use the checklist for short-answer items to identify the flaw(s), and rewrite the item so it remedies the flaw(s) you identified but creates no new flaws.
   a. ____ is the substance that helps plants turn light energy to food.
   b. The Johnstown Flood occurred during ____.
   c. The _____ is the major reason why ____ and _____ exhibit _____.
   d. San Francisco was named after ____.
   e. A kilogram is equivalent to ____.
   f. Was the population greater in 1941 or 1951?
3. Obtain a teacher's edition of a textbook (or other curricular materials) that covers the material for the teaching unit you selected for Exercise 3 of Chapter 6. Locate the completion and true-false items presented in the teacher's edition or textbook for this unit. Match those items to the learning targets included in the assessment blueprint you crafted for Exercise 3. To what extent do these items match the learning targets and the blueprint? What do your findings suggest about the way you should use the items the textbook gives you? About your need to craft items yourself? Prepare a short report and share your findings with others in this course.
4. Each of the following true-false items contains one or more flaws. For each item, use the checklist for true-false items to identify the flaw(s), and rewrite it, correcting the flaw(s) identified. Be sure your rewritten items do not exhibit new flaws.
   a. The two categories, plants and animals, are all that biologists need to classify every living thing. T F
   b. In the United States, it is warm in the winter. T F
   c. Editing assessment tasks is an important step in the assessment development process. T F
   d. The major problem in the world today is that too many people want more than their "fair share" of the Earth's resources. T F
   e. There were more teachers on strike in 1982 than in 1942, even though the employment rate was lower in 1942 than in 1982. T F
5. Write one true-false item in your teaching area(s) that assesses a student's use of each of the categories of propositions listed in Figure 8.2.

Multiple-Choice and Matching Exercises

KEY CONCEPTS

1. A multiple-choice item consists of one or more introductory sentences followed by a list of two or more suggested responses. The student must choose the correct answer.
2. Before writing a multiple choice item, consider your learning target and design both the question and answer choices to tap the level of understanding you want to measure.
3. Multiple-choice items have many advantanges, including the ability to assess a variety of learning targets efficiently. Multiple-choice items have been criticized because if they are used to excess, students do not have a chance to express their learning in their own words and may develop a "one right answer" mentality.
4. Do not use multiple-choice items if selecting from among alternatives does not reflect the learning target you are trying to assess.
5. Follow item-writing guidelines to create high-quality multiple-choice items.
6. Alternative varieties of multiple-choice items include greater-less-same, best-answer, experiment-interpretation, and statement-and-comment items.
7. A matching exercise presents a student with three things: (1) directions for matching, (2) a list of premises, and (3) a list of responses.
8. An advantage of matching exercises is their ability to assess student understanding of relationships. Criticisms include the fact that matching exercises are often used to test rote memorization, because these are the easiest kind of matching exercises to write.
9. Follow item-writing guidelines to create high-quality matching exercises.
10. Alternative varieties of matching exercises include masterlist and tabular formats.

IMPORTANT TERMS

"all of the above"

alternatives, choices, options

best-answer item

clueing, linking

context-dependent items, interpretive exercises, interpretive materials, linked items

correct-answer item

decontextualized knowledge

direct assessment, indirect assessment

directions for matching

distractor rationale taxonomy

distractors, foils

experiment-interpretation items

filler alternatives, deadwood alternatives

greater-less-same items

homogeneous alternatives, heterogeneous alternatives

homogeneous premises and responses

incomplete stem

keyed alternative, key, keyed answer

From Chapter 9 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

masterlist variety, classification variety, keylist variety

matching exercise (basic)

multiple-choice item

"none of the above"

overlapping alternatives

perfect matching

plausible distractors, functional alternatives

premise list

response list

statement-and-comment items

stem

tabular (matrix) items

tandem arrangement of options

clang associations, grammatical clues

window dressing

MULTIPLE-CHOICE ITEM FORMAT

A **multiple-choice item** consists of one or more introductory sentences followed by a list of two or more suggested responses. The student must choose the correct answer from among the responses you list. The following example illustrates this format:

Example

How many sides does a heptagon have?	} **Stem**
A Three	**Distractors**
B Five	
C Six	
*D Seven	} **Keyed alternative**

Note: Correct answers to multiple-choice items will be marked with an asterisk (*) throughout this book.

Stem

The **stem** is the part of the item that asks the question, sets the task a student must perform, or states the problem a student must solve. You write the stem so that a student understands what task to perform or what question to answer.

Alternatives

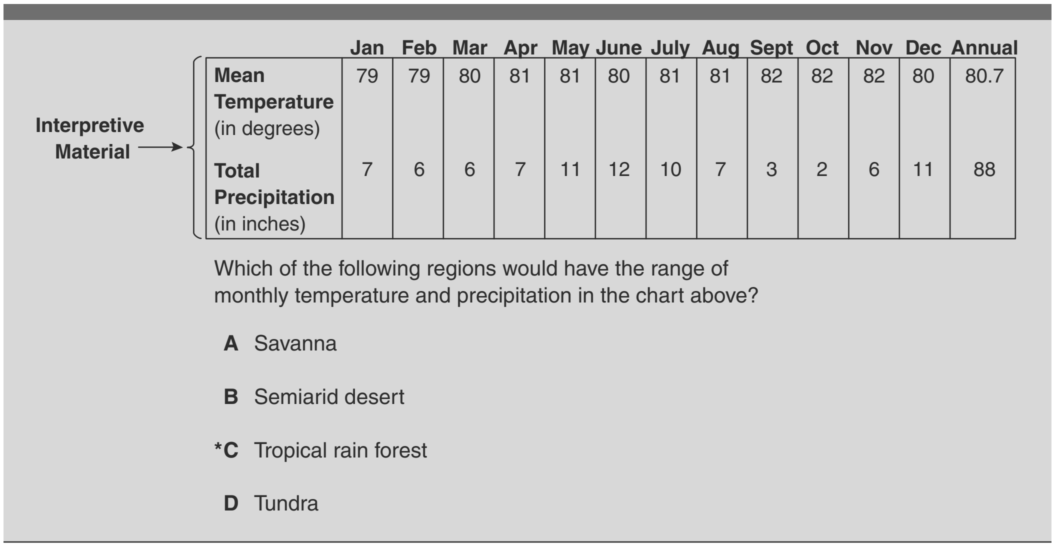
Teachers call the list of suggested responses by various names: **alternatives, choices,** and **options**. The alternatives should always be arranged in a meaningful way (logically, numerically, alphabetically, etc.). The chronological sequence in which events occur and the size of objects (large, medium, small) are examples of logical orders. If no logical or numerical order exists among them, the alternatives should be arranged in alphabetical order. In the preceding example, alternatives are in numerical order. The reason for this is that you do not want to establish a pattern that can clue the answer for students who do not know it. Second, following this rule saves the students time.

Keyed Alternative and Distractors

The alternative that is the correct or best answer to the question or problem you pose is called the **keyed answer, keyed alternative**, or simply the **key**. The remaining incorrect alternatives are called **distractors** or **foils**. The purpose of the latter is to present plausible (but incorrect) answers to the question or solutions to the problem in the stem. These foils should be plausible only to students who do not have the level of knowledge or understanding required by your learning target—those who haven't learned the material well enough. Conversely, the foils should not be plausible to students who have the degree of knowledge you desire.

Interpretive Material

In some cases, you may need to add information to make a question clearer or more authentic. You may wish to assess a learning target, for example, that requires students to apply their knowledge to data in a table or a graph, to a situation described in a paragraph, to an object, or to an event simulated by a picture. If adding this kind of information makes the stem more than one or two sentences long, then the information is placed in a section that comes before the stem. This information is called **interpretive material**, and the items that refer to it are called **context-dependent items**, **interpretive exercises**, or **linked items** (the items are "linked" to the interpretive material). Figure 9.1 illustrates this assessment technique. The table of weather data is the interpretive material. We give more elaborate suggestions for context-dependent items in Chapter 11.



Interpretive Material →

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Annual
Mean Temperature (in degrees)	79	79	80	81	81	80	81	81	82	82	82	80	80.7
Total Precipitation (in inches)	7	6	6	7	11	12	10	7	3	2	6	11	88

Which of the following regions would have the range of monthly temperature and precipitation in the chart above?

A Savanna

B Semiarid desert

***C** Tropical rain forest

D Tundra

FIGURE 9.1 Item with interpretive material.
Source: From *The Geography Learning of High-School Seniors* (p. 45), by R. Allen, N. Bettis, D. Kurfman, W. MacDonald, I. V. S. Mullis, and C. Salter (1990), Princeton, NJ: National Assessment of Educational Progress. Educational Testing Service. Reprinted by permission.

CONSIDERATIONS BEFORE WRITING ITEMS

Similarity of Distractors

Think of a student as being located at some point along a continuum of learning for a given learning target. You can construct a test item for students at a specific level of this learning continuum. The students who are at this level (or above it) should be able to answer the item correctly; others, lower on the continuum of learning, will not. Consider the following items:

Examples

1. In what year did the United States enter World War I?
   A 1776
   B 1812
   *C 1917
   D 1981
2. In what year did the United States enter World War I?
   A 1901
   *B 1917
   C 1941
   D 1950
3. In what year did the United States enter World War I?
   A 1913
   B 1915
   C 1916
   *D 1917

All three items ask the same question, but the specificity of knowledge that is required to answer that question increases from Item 1 to Item 3. In this example, you can easily see how the *alternatives operate to make the item easy or difficult*: The alternatives require the students to make finer distinctions among the dates. Some research supports the idea that similarity among the alternatives increases the difficulty of an item (Green, 1984). Although the example uses dates, similarities can also be the result of using certain words or concepts. Of course, manipulating the alternatives is not the only way to create more difficult items.

For which level of knowledge should an item be written? There is no general rule, but keep in mind these main points: the type of students, the level of instruction, the purpose for which you will use the assessment results, and the level of knowledge your students need to attain at this point in their educational development. Also consider the taxonomy thinking levels your test will assess. In effect, you need to decide, at least roughly, which level of proficiency is sufficient for each important learning target. Then construct test items that will allow you to distinguish students who lack sufficient proficiency from those who have acquired it. Or if you are trying to map students along a range of proficiencies (A, B, C, D, F, for example; or Basic, Proficient, Advanced; etc.), you should include items along the range of the continuum, so that each category of students will have some items that indicate the proficiency level.

Basic Purpose of Assessment Tasks

The preceding description represents an idealized situation. Seldom will your real assessment tasks separate students this neatly. Some less knowledgeable students probably will answer some tasks correctly, and other, more knowledgeable students will not. In general, though, keep in mind this principle:

> The basic purpose of an assessment task, whether or not it is a multiple-choice item, is to identify students who have attained a sufficient (or necessary) level of knowledge (skill, ability, or performance) of the learning target being assessed.

Varieties of Multiple-Choice Items

Teachers and professional test developers use several varieties of multiple-choice items. Some of these are shown in Figure 9.2.

Teachers usually find that the **correct-answer, best-answer,** incomplete-statement, and negative varieties in Figure 9.2 are the most useful. As you grow more skilled at evaluating students, you will find that you need to use several of these variations

FIGURE 9.2 Varieties of multiple-choice items.

A. The correct-answer variety

Who invented the sewing machine?

A. Fulton
*B. Howe
C. Singer
D. White
E. Whitney

B. The best-answer variety

What was the basic purpose of the Marshall Plan?

A. military defended western Europe
*B. reestablish business and industry in western Europe
C. settle United States' differences with Russia
D. directly help the hungry and homeless in Europe

C. The multiple-response variety

What factors are principally responsible for the clotting of blood?

A. contact of blood with a foreign substance
*B. contact of blood with injured tissue
C. oxidation of hemoglobin
D. presence of unchanged prothrombin

D. The incomplete-statement variety

Millions of dollars of corn, oats, wheat, and rye are destroyed annually in the United States by

A. mildews.
B. molds.
C. rusts.
*D. smuts.

E. The negative variety

Which of these is NOT true of viruses?

A. Viruses live only in plants and animals.
B. Viruses reproduce themselves.
*C. Viruses are composed of very large living cells.
D. Viruses can cause diseases.

F. The substitution variety

Passage to be read

Surely the forces of education should be fully utilized to acquaint youth with the real nature of the dangers to democracy*, for* no other place
1
offers *as good or better opportunities than* the school
2
for a *rational* consideration of the problems involved.
3

Items to be answered

1. *A. , for
B. . For
C. - for
D. no punctuation needed

2. A. As good or better opportunities than
B. as good opportunities or better than
C. as good opportunities as or better than
*D. better opportunities than

3. *A. rational
B. radical
C. reasonable
D. realistic

G. The incomplete-alternative variety[a]

An apple that has a sharp, pungent, but not disagreeably sour or bitter, taste is said to be (4)

A. p
B. q
*C. t
D. v
E. w

H. The combined response variety

In what order should these sentences be written in order to make a coherent paragraph?

a. A sharp distinction must be drawn between table manners and sporting manners.
b. This kind of handling of a spoon at the table, however, is likely to produce nothing more than an angry protest against squirting grapefruit juice about.
c. Thus, for example, a fly ball caught by an outfielder in baseball or a completed pass in football is a subject for applause.
d. Similarly, the dexterous handling of a spoon in golf to release a ball from a sand trap may win a championship match.
e. But a biscuit or a muffin tossed and caught at table produces scorn and reproach.

A. a, b, c, d, e
*B. a, c, e, d, b
C. a, e, c, d, b
D. b, e, d, c, a

[a] The numeral in parentheses indicates the number of letters in the correct answer (which in this case is "tart"). Using this number rules out borderline correct answers.

to obtain valid results. We will discuss some of these item varieties in this textbook. See Wesman (1971) for details about other varieties.

Direct Versus Indirect Assessment

A multiple-choice test can be a **direct assessment** of certain abilities. Well-written multiple-choice items, especially those requiring the use of interpretive materials, can help directly assess a student's ability to discriminate and make correct choices; to comprehend concepts, principles, and generalizations; to make judgments about and choices among various courses of action; to infer and reason; to compute; to interpret new data or new information; and to apply information and knowledge in structured situations.

Multiple-choice items are only **indirect assessments** of other important educational outcomes, such as the ability to recall (as opposed to recognize) information under minimal prompting conditions, to articulate explanations and give examples, to produce and express unique or original ideas, to solve problems that are not well structured, to organize personal thoughts, to display thought processes or patterns of reasoning, to work in groups, and to construct or build things. These are important abilities. Many of them can be assessed directly with other paper-and-pencil formats such as extended written assignments. Others require alternative assessment techniques such as observing a student over an extended period working alone or in a group; interviewing a student; or assessing a student's performance, product, or creation. These latter techniques are discussed in Chapter 12.

ADVANTAGES AND CRITICISMS OF MULTIPLE-CHOICE ITEMS

Advantages

The following are advantages of multiple-choice items:

1. *The multiple-choice format can be used to assess a greater variety of learning targets than other formats of response-choice items.* The various abilities were discussed in the preceding paragraphs.
2. *Multiple-choice items* (and other types of response-choice items) do not require students to write out and elaborate their answers and thus minimize the opportunity for less knowledgeable students to "bluff" or "dress up" their answers. Some consider this a disadvantage.
3. *Multiple-choice tests focus on reading and thinking.* They do not require students to use writing processes under examination conditions.
4. *Students have less chance to guess the correct answer to a multiple-choice item than a true-false item or a poorly constructed matching exercise.* The probability of a student blindly guessing the correct answer to a three-alternative item is 1/3; to a four-alternative item it is 1/4; and so on.
5. *The distractor a student chooses may give you diagnostic insight into difficulties the student is experiencing.* However, for distractors to work this way you must carefully craft them so they are attractive to students who make common errors or who hold common misconceptions. Note, too, that a single item is not a very reliable basis for a diagnosis. You will have to follow up to confirm your diagnosis.

Criticisms

Multiple choice items have been criticized on the following grounds. Most of these criticisms can apply to other types of assessments, as well.

1. *Students do not create or express their own ideas or solutions.* If you rely exclusively on multiple-choice testing, you will risk giving your students little or no opportunity to write about the topics in the subject they are learning.

2. *Poorly written multiple-choice items can be superficial, trivial, and limited to factual knowledge.* Of course, so can any poorly constructed assessment format. Gaining the knowledge and skill to overcome this criticism is the reason you are taking this course!

3. *Because usually only one option of an item is keyed as correct, brighter students may be penalized for not choosing it.* Brighter students may detect flaws in multiple-choice items due to ambiguities of wording, divergent viewpoints, or additional knowledge of the subject, whereas other students may not.

4. *Multiple-choice items tend to be based on "standardized," "vulgarized," or "approved" knowledge.* The problems students solve on multiple-choice items tend to be very structured and closed (having one correct answer). This gives the impression that all problems in a subject area have a single correct answer, which may encourage students to place too much faith in an authority figure's correctness or may misrepresent a subject area as having a

fixed and limited knowledge base. Further, if you use multiple-choice tests that fail to use items linked to realistic interpretive materials, tests do not have a real-world context. This is referred to as **decontextualized knowledge**. As a result, your tests may not assess whether students can use what they have learned in a meaningful and authentic context.

5. *Exclusive use of multiple-choice testing for important or high-stakes assessments may shape education in undesirable ways.* Those objecting to multiple-choice tests point out that the type of examination you use can shape the content and nature of instruction you deliver to students. If a high-stakes assessment's multiple-choice items focus on factual knowledge, teachers tend to use drill-and-practice techniques to prepare students for it. These strategies are less effective if the test contains multiple-choice items that assess using knowledge and applying higher-order thinking skills.

WHEN NOT TO USE MULTIPLE-CHOICE ITEMS

Definite "Don'ts"

Test items must be aligned with the student achievements you want them to assess. You would not, for example, substitute multiple-choice questions on English mechanics and grammar for actual samples of writing when your learning target calls for students to write. Nor would you use multiple-choice items when your main learning target requires students to organize their own ideas, develop their own logical arguments, express their own thoughts and feelings, or otherwise demonstrate their self-expression abilities.

When You Have a Choice

At times, you have a choice between using short-answer items for the entire test and using multiple-choice items for the entire test. On the surface, either format may seem appropriate. However, if most or all of the items in your test will assess students' simple recall, a short-answer test is preferred when:

1. Each of the items has only one correct answer and the correct answers are almost always a single word or number.
2. All of the items are computational problems calling for numerical answers.
3. Almost all of the items have only two possible plausible responses (e.g., yes vs. no, male vs. female, positive vs. negative).
4. The answer to each test item is short enough so that writing the answers doesn't take the student any longer than marking the answer to multiple-choice questions on an answer sheet.

When any one of these situations exists, it will be difficult for you to write good multiple-choice items requiring students to demonstrate the required degree of recollection and computation. Further, as in Situations 2 and 4, sometimes there is no advantage to the multiple-choice format over the more direct short-answer format.

Exceptions

There are exceptions to these suggestions, of course. When you need to assess a large number of students over large areas of content, and when you have readily available machine scoring, multiple-choice items may be the only practical assessment. Or when you already have a test that includes lots of good multiple-choice items but only one or two of those items fit one of the four situations previously listed, it is more efficient to use the multiple-choice format for these items.

If your students will be administered a standardized achievement test (either by your school district or by the state), it will be to their advantage to have experience answering multiple-choice items. In fact, some educational measurement specialists argue that in such instances you would be remiss if you did not give your students practice in taking multiple-choice tests. Therefore, you may wish to use multiple-choice items for at least some parts of your assessments to give students appropriate practice, even though one of the four situations exists.

Classroom assessment generally benefits from a mixture of assessment formats. Create each task to best assess the respective learning targets. The validity of your results, rather than your own convenience, should be your first priority.

Other "Don'ts" and Exceptions

Some writers advise against including multiple-choice items when the test will be used only once, or when there are few students. It is easier to formulate short-answer questions than to write good multiple-choice items, and scoring will not be time-consuming when there are few students. However, even when the number of students is small, if you plan to teach the same subject at the same level in subsequent years, it is usually worthwhile to develop a “pool” of multiple-choice items over time. You can then select items from this pool for future tests.

CREATING BASIC MULTIPLE-CHOICE ITEMS

Five Basic Skills of the Craft

You will create useful multiple-choice items if you learn how to do five things: (1) focus items to assess specific learning targets; (2) prepare the stem as a question or problem to be solved; (3) write a concise, correct alternative; (4) write distractors that are plausible; and (5) edit the item to remove irrelevant clues to the correct answer. First-draft multiple-choice items should not be put on a test until they are edited and polished. Editing items is a necessary step, even for the most experienced item writers. This section presents several item-writing guidelines for improving items in this editorial stage.

Our suggestions for crafting multiple-choice items are organized into three groups: suggestions for the stem, suggestions for the distractors, and suggestions for the correct alternative. Suggestions for improving the quality of the stem portion of multiple-choice items are summarized in Figure 9.3 and discussed in more detail in the following section.

Crafting the Stem of the Item

Direct Question Asked or Implied After reading the stem, a student should understand the main intent of the item—what type of response you expect. The stem should ask a direct question or should clearly formulate a problem for the student to solve.

Incomplete sentences sometimes make good stems, but experience and research (Haladyna, Downing, & Rodriguez, 2002) indicate that item writers usually produce better items when they phrase the stem as a direct question. The reason is probably that when a teacher does not ask a direct question, the student must mentally rephrase the stem as a question appropriate to the alternatives presented. This increases the cognitive complexity of the student’s task, perhaps beyond what you may intend. When an incomplete sentence is used, a question is implied, of course. Older and brighter students are sometimes able to do this rephrasing without difficulty. However, younger, more average students, and perhaps those experiencing some learning difficulties, may find that this extra process increases their difficulty in expressing what they know.

A simple way to check for this flaw is to cover the alternatives with your hand. Then, read the stem. On the basis of that stem alone, can you determine what is expected of the student? If not, the stem is incomplete and you should rewrite it. The example that follows shows how an item is improved by rephrasing the incomplete stem as a question:

Example

Poor: Incomplete stem

1. W. E. B. DuBois
   *A actively pressed for complete political participation and full rights for African Americans.

FIGURE 9.3 **Suggestions for improving the quality of the stems of multiple-choice items.**

To do	To avoid
1. If possible, write as a direct question. 2. If an incomplete sentence is used, be sure a. it implies a direct question. b. the alternatives come at the end (rather than in the middle) of the sentence. 3. Control the wording so that vocabulary and sentence structure are at a relatively low and nontechnical level. 4. In items testing definitions, place the word or term in the stem and use definitions or descriptions as alternatives.	1. Avoid extraneous, superfluous, and nonfunctioning words and phrases that are mere “window dressing.” 2. Avoid (or use sparingly) negatively worded items. 3. Avoid phrasing the item so that the personal opinion of the examinee is an option. 4. Avoid textbook wording and “textbookish” or stereotyped phraseology. 5. Avoid “cluing” and “linking” items (i.e., having the correct answer to one item be clued or linked to the correctness of the answer to a previous item).

B taught that the immediate need was for African Americans to raise their economic status by learning trades and crafts.
C emphasized helping African Americans through the National Urban League.
D founded the Association for the Study of Negro Life and History.

Better: Asks a question

2. Which of the following comes closest to expressing W. E. B. DuBois's ideas about priorities of activities of African Americans during the early 20th century?
A African Americans should first improve their economic condition before becoming fully involved in politics.
B African Americans should postpone the fight for equal access to higher education until their majority acquire salable trade skills.
C African Americans should withdraw from white society to form a separate state in which they have complete political and economic control.
*D African Americans should become active, seeking out complete citizenship and full political participation immediately.

Question 1 is poor because the stem does not set a task or ask a question. (Cover the alternative. What task or problem does the stem set?) The student must read the entire item and infer that the teacher must be trying to find out something about W. E. B. DuBois's ideas. The student may very well know DuBois's ideas, but if the student makes the wrong inference about the teacher's intent, the student may answer the item incorrectly. Question 2 is better because the intent of the item is clear after the student reads the stem.

Put Alternatives at the End This rule is similar to the rule for completion items to place the blank at the end of the incomplete sentence (Chapter 8). The following example shows how an item is improved by listing alternatives at the end of the stem:

Example

Poor: Options in the middle of the stem

1. Before the Civil War, the South's
*A emphasis on staple-crop production
B lack of suitable supply of raw materials
C short supply of personnel capable of operating the necessary machinery
was one of the major reasons manufacturing developed more slowly than it did in the North.

Better: Options put at the end

2. Before the Civil War, why did manufacturing develop more slowly in the South than in the North?
*A The South emphasized staple-crop production.
B The South lacked a suitable supply of raw materials.
C The South had a short supply of people capable of operating the necessary machinery.

Control Vocabulary and Sentence Structure When testing for subject-matter learning, make sure you phrase the item at a level suitable for the students. You don't want long sentences, difficult vocabulary, and unnecessarily complex sentence structures to interfere with students' ability to answer the item. This may be especially true when you have students with disabilities mainstreamed in your class. For example, students with hearing disabilities frequently have relatively large language and vocabulary deficits. These students may very well have acquired the specific knowledge, concept, or principle you are assessing, but the way you phrase an item may interfere with their ability to demonstrate this knowledge. The example that follows illustrates how a simple information item can be complicated by uncontrolled language. The item is improved by making it more concise.

Example

Poor: Unnecessary wordiness and complexity

1. Given the present-day utilization of the automobile in urban settings, which of the following represents an important contribution of Garrett A. Morgan's genius?
A automobile safety belts
B crosswalk markers
*C traffic lights
D vulcanized rubber tires

Better: More concise

2. Which of the following did Garrett A. Morgan invent?
A automobile safety belts
B crosswalk markers
*C traffic lights
D vulcanized rubber tires

Avoid "Window Dressing" Item 1 in the preceding example demonstrates how extraneous wording can unnecessarily complicate an item. Less obvious is the use of words that tend to "dress up" a stem to make it sound as though it is testing something of practical importance. Often such **window dressing** creeps into an item when you are struggling to measure higher-level cognitive abilities, such as applications. Window dressing makes an item appear to measure applications, when it does not. The next example shows how window dressing makes an item more difficult, less discriminating, less reliable, and less valid. The item is improved by eliminating the window dressing.

Example

Poor: Window dressing

1. There are 10 preservice teachers in the Department of Education who recently registered for the college-sponsored weight loss program. At the beginning of the program each was weighed, and the 10 had a mean weight of 139.4 pounds. Suppose there were but three men in this group, and that their mean weight was 180 pounds. What was the mean weight of the women at the beginning of the program?
   A 115.0 pounds
   *B 122.0 pounds
   C 140.0 pounds
   D 159.7 pounds

Better: More concise

2. Ten persons have a mean weight of 139.4 pounds. The mean weight of three of them is 180 pounds. What is the mean weight of the remaining seven persons?
   A 115.0 pounds
   *B 122.0 pounds
   C 140.0 pounds
   D 159.7 pounds

Every word used in an item should have a purpose. Sometimes names, places, and other "facts" about a situation are necessary pieces of information: They can give the student the basis for determining the correct answer. The following example shows an acceptable inclusion of facts in an item stem:

Example

A company owns a fleet of cars for which it pays all fuel expenses. Three readily available types of gasoline were tested to see which type was giving better mileage. The results are shown below in miles per gallon.

	Mean	*Median*
Type A	19.1	18.5
Type B	18.5	19.1
Type C	18.8	18.9

1. Assuming they all cost the same, which type of gasoline should the company use?
   *A Type A
   B Type B
   C Type C

Avoid Negatively Worded Stems Phrase items positively if possible. Negatively worded stems, such as "which of the following is not . . . ," tend to confuse students, especially the younger or less careful ones. Even well-prepared students often overlook the *not* in an examination question. Positively worded items are easier for students than the corresponding negatively worded items (Haladyna & Downing, 1989a; Haladyna et al., 2002). The following example shows how to improve an item by using positive wording:

Example

Poor: Negatively phrased stem

1. Sometimes a teacher finds it necessary to use a mild form of punishment. When this occurs, which of the following should not happen?
   A Children should not believe all of their behavior is bad.
   B Children should understand the reason(s) why they are being punished.
   *C Children should understand that the teacher, not them, controls when the punishment will end.

Better: Positively phrased stem

2. Sometimes a teacher finds it necessary to use a mild form of punishment. When this occurs, it is important that the children understand
   A that it may be a long time before happy times return to the classroom.
   *B the reason(s) why they are being punished.
   C that the teacher, not the children, controls when their punishment will end.

If negatively phrased items must be used, use the negative word only in the stem or only in an option (not both), and either underline the negative word or place it in CAPITAL LETTERS.

Avoid Grading Personal Opinions Do not ask for students' personal opinions in the context of a multiple-choice test in which the students need to select one option as best or correct. Everyone is entitled to an opinion. If you ask for students' opinions in a multiple-choice item, every option could be correct. The following example illustrates this point:

Example

Poor: Makes the correct answer a matter of personal opinion

1. Which of the following men contributed most toward the improvement of the self-confidence of African Americans?
   A W. E. B. DuBois
   B Eugene K. Jones
   C Booker T. Washington
   D Carter G. Woodson

There is no single correct answer to the preceding question, because each man's contributions can be judged and evaluated in different ways. The question could form the basis for an extended-response essay or a term paper in which the students support their opinions with evidence and logical argument. In that case, do not grade the opinions or the positions taken. Rather, evaluate the way the students use the evidence to support their opinions.

Avoid Textbookish Wording As with true-false items, when you copy sentences verbatim from the text you end up with a poor item because (a) frequently, a sentence loses its meaning when you take it out of context, (b) you encourage rote memory of textbook material instead of comprehension, (c) you are likely to produce awkwardly worded items with implausible distractors, and (d) learners who have only a superficial understanding of the underlying concept or principle may obtain clues to the correct answer by simply recalling the textbook phrasing. Use a new, perhaps less familiar, wording of the stem and correct option to test a deeper comprehension of a concept or principle. Avoid textbookish phrasing. Even though you paraphrase, you may word the item so it reads very much like the textbook. The procedure we discussed earlier in Chapter 8—stating main ideas of textbook passages in your own words and rephrasing these as questions—is a practical one for avoiding textbookish phrasing. Here is an example:

Example

Poor: Uses textbookish phrasing

1. The annual incomes of five employees are $8,000, $8,000, $10,000, $11,000, and $25,000, respectively. Which index should be used to summarize the typical employee's income?
   A mean
   *B median
   C mode

Better: Novel situation for students

A teacher keeps a record of how long it takes students to complete the 50-question final exam. The mean time was 46 minutes and the median time was 20 minutes. The teacher used this information to set the next exam's time at 20 minutes. The teacher reasoned that these data demonstrate that the typical student could complete the test in that time.

2. In all likelihood, this time limit is
   A just about right.
   *B too short.
   C too long.

The first item is weak because most introductory statistics books associate the term *typical* with median and often use income examples to illustrate the application of the median. By knowing these superficial facts, the student can mark "B" without demonstrating an in-depth understanding of this statistical index. The better item, Question 2, assesses a different learning target. It is better because it presents a novel situation and requires an application of the concept.

When testing older students, you may find it helpful to use stereotyped phraseology, certain "pat phrases," and verbal associations to make distractors plausible to students lacking the required degree of knowledge. These phrases may be put into the stem or into the distractors. Item 2, although it is a bit wordy and places a premium on reading, does just this. A student who interprets the correctness of using a particular statistical index only on the basis of the verbal association of *typical* with median will not answer the item correctly. Such a student will fail to notice that if the teacher set the time limit for the test at 20 minutes, only half of the class will have enough time to complete it. Surely this is inappropriate for a classroom test.

Create Independent Items With the possible exception of context-dependent items (see Chapter 11), each item should assess a distinct performance, and the correct answer to an item should not be clued by another item. Two flaws to avoid are linking and clueing. **Linking** means that the answer to one or more items depends on obtaining the correct answer to a previous item. Linked items frequently result in a double penalty for an incorrect answer, as when a computational result from one item is required to answer a subsequent item. **Clueing** means that a hint to the correct answer to one item is found in the contents of another item in the test. In the next example, Questions 1 and 2 illustrate linked items:

Example

Preceding item

1. The perimeter of a rectangle is 350 centimeters. The length of the rectangle is 3 centimeters longer than the width. What is the width?
   A 18.7 cm.
   *B 86.0 cm.
   C 89.0 cm.
   D 116.7 cm.

Poor subsequent item: Linked to Item 1

2. What is the area of the rectangle described in Question 1?
   A 1,050 sq. cm.
   B 7,396 sq. cm.
   *C 7,654 sq. cm.
   D 8,188 sq. cm.

Better subsequent item: Independent of Item 1

3. The width of a rectangle is 4 centimeters and the length is 3 centimeters. What is the area?
   A 9 sq. cm.
   *B 12 sq. cm.
   C 16 sq. cm.
   D 17 sq. cm.

The "preceding item" is primarily computational. The poor subsequent item (Item 2) is linked to it. A student could make an incorrect computation in Item 1, obtaining 89.0, for example. Having already made a mistake in Item 1, the student also would get Item 2 wrong, because 89 × (89 + 3) = 8,188 is keyed as the wrong answer. One solution to the problem is shown in the better subsequent item: You present a new numerical value for the student to use, which is independent of the preceding item. Thus, Item 3 is not linked to Item 1.

Of course, items may provide clues to other items even though they are not linked. Review all the items in your test to see if any item suggests an answer to other items.

Definitions Go in the Alternatives Teachers frequently assess whether students know the meaning of special terms or vocabulary words. Multiple-choice items are often used for this purpose. A common flaw, however, is to put the definition of the term in the stem and to use a list of words as alternatives. The flaw with this approach is that it increases the likelihood that students will get the answer correct by using only superficial knowledge of the definition being assessed. Students can obtain the correct answer by knowing only that the words in the definition "look like" (seem similar to) a word in the alternatives. To assess whether students have in-depth knowledge of a term, put the term in the stem and write various definitions in the alternatives. The item can be made easier or harder depending on how similar the alternatives are. The following example shows how to improve a definition item by putting the term in the stem and using different definitions as alternatives:

Example

Poor: Definition in the stem

1. The increase in length per unit of length of a metal rod for each degree rise in temperature (Centigrade) is known as the
   *A coefficient of linear expansion of the metal.
   B elasticity of the metal.
   C specific heat of the metal.
   D surface tension of the metal.

Better: Definitions in the alternatives

2. What is the *coefficient of linear expansion* of a metal rod?
   A the increase in length of the rod when its temperature is raised 1°C
   *B the increase in length when the temperature is raised 1°C divided by the total length of the rod at its original temperature
   C the ratio of its length at 100 to its length at 0°C
   D the rise in temperature (degrees Centigrade) that is necessary to cause the length of the rod to expand 1 percent

Source: Adapted from "The Construction of Tests," by E. F. Lindquist and C. R. Mann, in *The Construction and Use of Achievement Examinations: A Manual for Secondary School Teachers* (pp. 145–146), by H. E. Hawkes, E. F. Lindquist, and C. R. Mann (Eds.), 1936, Boston: Houghton Mifflin.

Crafting Alternatives or Foils

The alternatives of a multiple-choice item present choices to the students. All of the choices must be appropriate to the stem. If they are not, they may be confusing to knowledgeable students or may be easily eliminated by less knowledgeable ones. When all alternatives are appropriate to the stem, the item functions better as a complete unit. Suggestions for improving the quality of the alternatives are summarized in Figure 9.4 and discussed in more detail in the following text.

Plausible and Functional Alternatives

Many of the suggestions that follow will help you craft plausible distractors. **Plausible distractors** are incorrect alternatives that appear to be correct to students who have not mastered the assessed learning target. To make distractors plausible, base them on errors students commonly make, such as computational errors, conceptual errors, or errors resulting from faulty common knowledge. In this way, your analysis of students' responses could help you identify their specific difficulties.

Figure 9.4 calls for using from three to five functional alternatives. A **functional alternative** serves the purpose for which it is written. This means that an alternative which is a distractor attracts at least one of the students who do not have the degree of knowledge that you expect of all students. Also, an alternative that is the keyed answer is functional if all students who do have the degree of knowledge you expect select it.

For example, an item may have five alternatives. If even the most superficial learner easily eliminates two of the distractors, however, only the remaining three are seriously considered plausible answers. In reality, then, the item has only three *functional* alternatives. For practical purposes you may as well delete the two nonfunctional alternatives. Nonfunctional distractors are called **"deadwood"** or **filler alternatives**.

Teachers sometimes ask if each multiple-choice item should have the same number of alternatives and, if so, how many there should be. Assessment specialists have long recognized that there is no virtue in having the same number of alternatives for each item. This is especially true for classroom assessment. Research supports the rule that you should write as many functional distractors as is feasible; as Haladyna and Downing (1989b) point out, "The key to distractor development is not the *number* of distractors but the *quality* of distractors" (p. 59).

If you can write three to five functional alternatives, then the item is more likely to distinguish those who have the desired degree of knowledge from those who do not. Research suggests that having three functional alternatives is best, on balance (Rodriguez, 2005). The more alternatives you try to write, the harder it will be to make them functional. As a rule of thumb, strive to write three functional alternatives for most purposes, and use up to five functional alternatives if there is a justification for each (for example, if each distractor exemplifies a common kind of error). Don't waste your time trying to create the same number of alternatives for each item if by so doing you are creating nonfunctional fillers or deadwood. If a separate answer sheet is used for machine scoring, check the maximum spaces allowed per item and adjust the maximum number of alternatives accordingly.

Homogeneous Alternatives

Lack of homogeneity is a primary reason why distractors do not function. An item is said to have **homogeneous**

FIGURE 9.4 Suggestions for improving the alternatives of multiple-choice items.

To do	To avoid
1. In general strive for creating three to five functional alternatives.	1. Avoid overlapping alternatives.
2. All alternatives should be homogeneous and appropriate to the stem.	2. Avoid making the alternatives a collection of true-false items.
3. Put repeated words and phrases in the stem.	3. Avoid using "not given," "none of the above," etc. as an alternative in *best-answer* type of items (use only with *correct-answer* variety).
4. Use consistent and correct punctuation in relation to the stem.	4. Avoid using "all of the above": limit its use to the *correct-answer* variety.
5. Arrange alternatives in a list format rather than in tandem.	5. Avoid using verbal clues in the alternatives.
6. Arrange alternatives in a logical or meaningful order.	6. Avoid using technical terms, unknown words or names and "silly" terms or names as distractors.
7. All distractors should be grammatically correct with respect to the stem.	7. Avoid making it harder to eliminate a distractor than to choose the keyed alternative.

alternatives when each alternative belongs to the same set of "things" *and* each alternative is appropriate to the question asked or problem posed by the stem. For example, if the stem asks students to identify the name of someone who invented a particular machine, then each alternative should be a name and each name should be an inventor to be appropriate to the stem.

An item has **heterogeneous alternatives** when one or more of its alternatives do not belong to the same set of things. The following example shows how to improve an item by making its alternatives homogeneous:

Example

Poor: Heterogeneous alternatives that do not belong to the same category

1. What is the official state bird of Pennsylvania?
   A mountain laurel
   B Philadelphia
   *C ruffed grouse
   D Susquehanna River

Better: Homogeneous alternatives that belong to the same category

2. What is the official state bird of Pennsylvania?
   A goldfinch
   B robin
   *C ruffed grouse
   D wild turkey

You may also adjust the degree of homogeneity to control the difficulty of an item. The World War I items in the "Considerations Before Writing Items" section illustrated this point previously. Whether alternatives are perceived to be homogeneous by the students you are assessing depends on their level of educational development. Item 2 in the preceding example could be made more homogeneous (and more difficult) by using as alternatives the scientific names of several different species of grouse, for example. The World War I items illustrate this point, as well: The alternatives in Item 1 in that section may appear homogeneous to less knowledgeable, younger students, but they will likely appear to be quite heterogeneous alternatives to knowledgeable, older students.

Put Repeated Words in the Stem In general, it is better to put into the stem words or phrases that are repeated in each alternative. A more complete stem reduces the amount of reading required of the students and makes the task clearer to the student. To accomplish this, you may find it necessary to rephrase the stem to focus it on the critical point of the learning target. The next example shows how to improve an item by eliminating words that are repeated in each alternative:

Example

Poor: Words repeated in each alternative

1. Which of the following is the best definition of *seismograph*?
   A an apparatus for measuring sound waves
   B an apparatus for measuring heat waves
   *C an apparatus for measuring earthquake waves
   D an apparatus for measuring ocean waves

Better: Stem is more focused and repeated words are incorporated into the stem

2. What type of waves does a *seismograph* measure?
   *A earthquake waves
   B heat waves
   C ocean waves
   D sound waves

Consistent, Correct Punctuation If the stem asks a direct question (i.e., it ends in a question mark), the options can be either (a) complete sentences; (b) single words, terms, names, or phrases; or (c) other incomplete sentences. Complete sentences begin with a capital letter and end with an appropriate punctuation mark; do not use a semicolon or other inappropriate terminal punctuation. If the options are single words or incomplete sentences, do not use terminal punctuation. However, use a consistent rule for capitalizing the initial word in each option: Throughout the test, either capitalize *all* initial words or capitalize *no* initial word (except a proper noun, of course).

An **incomplete stem** contains an incomplete sentence that the student must complete by choosing the correct alternative. In this case, choose alternatives to complete the sentence that would be plausible for students who have not mastered the learning target. When writing this type of item, begin each alternative with a lowercase letter (unless an alternative's initial word is a proper noun) and end it with the appropriate terminal punctuation.

There are exceptions to these rules, of course, as when the purpose of an item is to assess knowledge of grammar rules. The next item illustrates this exception to the rule of using consistent punctuation:

Example

Choose the phrases that correctly complete the sentence.

1. Julia became very frightened and shouted,
   A "Please save me."
   B "Please save me?"
   *C "Please save me!"

Arrangement of the Alternatives Alternatives are less confusing and easier to read when they are arranged one below the other in list form rather than in a **arrangement of options tandem** (beside one another). The following item shows a poor tandem arrangement of the alternatives:

Example

Poor: Tandem arrangement of alternatives

1. The angles of a triangle measure 80°, 50°, and 50°. What type of triangle is it?
   A Equilateral triangle
   *B Isosceles triangle
   C Obtuse triangle
   D Right triangle

Alternatives should be arranged in meaningful order, such as order of magnitude or size, degree to which they reflect a given quality, chronologically, or alphabetically. Such arrangements make locating the correct answer easier for the knowledgeable student, reduce reading and search time, and lessen the chance of careless errors. The next examples show acceptable arrangements of alternatives:

Examples

Alphabetical arrangement of alternatives

1. Which of the following is made from the shells of tiny animals?
   *A chalk
   B clay
   C shale

Numerical arrangement of alternatives

2. A student's percentile rank is 4. What is the stanine corresponding to this percentile rank?
   A 4
   B 3
   C 2
   *D 1

Grammatically Correct Relationship to the Stem Items that contain grammatical clues to the correct answer are easier and less reliable than items without such clues (Haladyna & Downing, 1989a). Don't clue the correct answer or permit distractors to be eliminated on superficial bases. Examples of inappropriate **grammatical clues** include lack of subject-verb agreement, inappropriate indefinite article, and singular/plural confusion. Below are examples of improving items by eliminating grammatical clues to the correct answers:

Examples

Poor: The definite article "a" at the end of the stem and plural usage of "angles" in the alternatives clue the correct answer.

1. A 90° angle is called a
   A acute angles.
   B obtuse angles.
   *C right angle.

Better: Writing the stem as a direct question eliminates the grammatical clue.

2. What are 90° angles called?
   A acute angles
   B obtuse angles
   *C right angles

Poor: Only one alternative uses the conjunction in a grammatically correct relationship to the stem.

3. Green plants may lose their color when
   A are forming flowers.
   *B grown in the dark.
   C are placed in strong light.
   D temperature drops.

Better: Ask a direct question to focus the item.

4. When may green plants lose their color?
   A when they form flowers
   *B when they are grown in the dark
   C when they are placed in strong light
   D when the surrounding temperatures drop

The indefinite article *a* in Item 1 gives the student the clue that Alternative C is correct. The other two alternatives begin with vowels and thus require the indefinite article *an*. In Item 3, the conjunction *when* is appropriate only to the phrasing of Alternative B.

Overlapping Alternatives Each alternative should be distinct and not a logical subset of another alternative. Alternatives that include some or all of one another are called **overlapping alternatives**. If you write an item containing overlapping alternatives, you give the less knowledgeable, but testwise, student clues to the correct answer. Several examples of improving items written with overlapping alternatives follow:

Examples

Poor: All alternatives have essentially the same meaning.

1. Why is there a shortage of water in the lower basin of the Colorado River?
   A The hot sun almost always shines.
   B There is a wide, hot desert.
   C The temperatures are very hot.
   *D All of the above are reasons why.

Better: Each alternative has a distinct meaning.

2. Why is there a shortage of water in the lower basin of the Colorado River?
   *A There is low rainfall and few tributaries at that region.
   B The desert soaks up water quickly.
   C A dam in the upper part made the lower part dry up.

In Item 1, Options A, B, and C essentially say the same thing: A testwise student, recognizing this overlap, would likely choose Option D even if the student knew nothing about the need for water in the lower Colorado River basin.

Avoid a Collection of True-False Alternatives A frequent cause of this type of flaw is that the teacher did not have in mind a clear problem or question when creating the item. Here is an example of improving an item by refocusing the collection of true-false alternatives:

Example

Poor: Alternatives are an unfocused collection of true-false statements.

1. A *linear function* is
   *A completely determined if we know two points.
   B completely determined if we know one point.
   C unrelated to the point-slope formula.
   D the same as the y intercept.

Better: The stem focuses on a problem.

2. In which of the following situations would it be possible to write the *equation for a linear function*?
   *A We know the line passes through the points (3,5) and (4,6).
   B We know the slope is 1.
   C We know the y intercept is (0,2).

In Item 1, it is difficult to identify any single question to which a student must respond. All options are related only by the fact that they could begin with the phrase, "A linear function is." Options B, C, and D, when used with that phrase, become false statements. This item is unfocused because it really embeds three ideas: how two points determine a line, the definition of the point-slope formula, and the definition of y intercept. Only one of these ideas should be selected and used as a basis for a revised item, as is done with Item 2. Or, rewrite the item as a multiple true-false item (see Chapter 8).

Avoid "None of the Above" Research on the phrase **"none of the above"** as an option in multiple-choice items indicates that it results in less reliable, more difficult items (Haladyna et al., 2002). Therefore, be very cautious when using this phrase as an option. This option should never be used with the best-answer variety (see the example given earlier in the chapter) of multiple-choice items. The very nature of a best-answer question requires that all of the options are to some degree incorrect, but one of them is "best." It seems illogical to require students to choose "none of the above" under these conditions.

It does make sense, however, to use "none of the above" with some correct-answer questions, when students look for one option that is completely correct. In areas such as arithmetic, certain English mechanics, spelling, and the like, a single, completely correct answer can be definitely established and defended. Some assessment experts recommend using "none of the above" only when students are more likely to solve a problem first before looking at the options, as opposed to searching through the distractors before proceeding with the solution to the problem.

Two special problems associated with using "none of the above" are (1) students may not believe that this choice can be correct and, therefore, they do not think it is plausible; and (2) students who choose it may be given credit when their thinking is incorrect. To avoid the first problem, use "none of these" as the correct answer to a few easy items near the beginning of the test. Students will then seriously consider "none of the above" as a possible correct answer for the remainder of the test. It may then be used as either a correct or incorrect answer later in the test. The second problem is handled by using "none of the above" as a *correct answer* in an item when the distractors encompass most of the wrong answers that can be expected (Item 1 below), or using it as a *distractor* for items in which most of the probable wrong answers cannot be incorporated into the distractors (Item 2).

Example

Acceptable use of "none of the above": As a correct answer

1. What is the difference?

```
 106        A  81
 −21        B  89
   ?        C  101
           *D  None of the above
```

Acceptable use of "none of the above": As a plausible distractor

2. What is the sum?

```
 46        *A  141
 47         B  161
 48         C  171
 ?          D  None of the above
```

More than likely, however, items such as 1 and 2 would be better as completion (short-answer) items than as multiple-choice items. If you used completion items, you would be able to check the students' wrong answers to determine why they responded incorrectly; then you could provide students with remediation.

Two final comments on this point: Avoid using "none of the above" as a filler to increase the number of distractors. Remember that distractors must be plausible. Second, as an option, "none of the above" is probably more confusing to younger students than to older ones.

Avoid "All of the Above" Research on the use of **"all of the above"** is inconclusive (Haladyna et al., 2002). This option, if used at all, should be limited to correct-answer varieties of multiple-choice items. It cannot be used with best-answer varieties because "all of the options" cannot simultaneously be best. Two further difficulties arise: (1) Students who know that one option is correct may simply choose it and inadvertently go on to the next item without reviewing the remaining options, and (2) students who know that two out of four options are correct can choose "all of the above" without knowing the correctness of the third option. The first difficulty can be reduced to some extent by making the first choice in the list read "all of the following are correct." However, this wording can also confuse elementary and junior high students. Generally, the recommendation is to avoid using "all of the above." Rewrite items with multiple answers as two or more items and avoid these problems. Alternately, rewrite the item as a multiple true-false item (see Chapter 8).

Avoid Verbal Clues Failure to follow this rule makes items easier and lowers the test reliability (Haladyna & Downing, 1989b). Verbal clues include using overlapping alternatives, silly or absurd distractors, **clang associations** (i.e., soundalike words) or other associations between words in the stem and in the correct alternatives, repetition or resemblance between the correct alternative and the stem, and specific determiners. Verbal clues in the alternative frequently lead the less knowledgeable but verbally able student to the correct answer. An example follows:

Example

Poor: Answered by association of words in stem and in correct answer

Which government agency is most concerned with our nation's agricultural policies?

*A Department of Agriculture
B Department of Education
C Department of the Interior
D Department of Labor

The item above, for example, uses *agriculture* in both stem and alternative. This creates a "Who is buried in Grant's tomb?" type of question.

Specific determiners are words that overqualify a statement so that it is always true or always false. We saw how these operated with true-false items in Chapter 8; they can occur in multiple-choice items as well. Students can eliminate options that state something "always" or "never" happens, for example, without really thinking about their content.

Avoid Technical and Unfamiliar Wording Teachers writing multiple-choice items sometimes use highly technical or unfamiliar words as distractors. This results in students needing more ability to reject the wrong answer than to choose the correct answer. Some studies indicate, however, that students view options containing unfamiliar technical words as less plausible, thereby making such alternatives nonfunctional.

Do Not Make a Distractor Too Plausible Incorrect alternatives sometimes may be made so plausible that generally good students get the item wrong, whereas less able students respond correctly. (Such items are said to be *negatively discriminating*; see Chapter 13.) The good students' knowledge, though perhaps normally sufficient for selection of the correct answer when embedded in another context, may be insufficient for rejection of all the distractors in a particular item. The history items in Figure 9.5 illustrate how students' insufficient knowledge and wrong learning can result in poorly functioning items.

Writing the Correct Alternative

You should word the correct alternative so that students *without* the requisite knowledge are *not clued* as to the correct answer and those students *with* the requisite knowledge *are able* to select the correct answer.

1. *In general, there should be only one correct or best answer to a multiple-choice item.* It is possible to write items that have more than one correct alternative. However, such items may not be as valid as you intend, especially with elementary and junior high school students. Students may, for example, mark the first correct alternative they encounter and skip to the next item without considering all of the alternatives. Some beginning item writers attempt to compensate for this behavior by using the combined response variety of multiple-choice items (see the example of this type given earlier in this chapter) or by using "all of the above." This usually results in poorer-quality items.

FIGURE 9.5 **Effects of insufficient or incorrect learning.**

The Effect of Insufficient Learning or Understanding

The failure of an item to function because of insufficient or wrong learning may be something beyond the control of the test constructor. . . .

What was one of the most important immediate results of the War of 1812?

1. The introduction of a period of intense sectionalism (39%)
2. The destruction of the United States Bank (7%)
3. The defeat of the Jeffersonian Party (7%)
4. The final collapse of the Federalist Party (4% omitted the item) (43%)

The correct response is Option 4. Nevertheless, the pupils who selected the first and incorrect response were, on the average, superior in general achievement to those who selected the correct response (4). The pupils selecting the first and incorrect response apparently did so because of positive but insufficient learning. They did know that a period of intense sectionalism set in before the middle of the century, and therefore chose the first response. Apparently they did not know, or failed to recall, that a short period of intense nationalism was an immediate result of the Second War with Great Britain, and that this war, therefore, could not be considered as "introducing" an era of sectional strife. Other pupils, with less knowledge in general, were able to select the correct response because they were not attracted to the first response by a certain knowledge that intense sectionalism did develop in the 19th century. (It should be noted, however, that for an abler group of pupils, capable of making the judgment called for, this same item might have shown a high positive index of discrimination.)

The Effect of Wrong Learning

Wrong learning, as well as insufficient learning, on the part of pupils for whom the test is intended may cause an item in that test to show a negative index of discrimination.

In the second half of the 15th century the Portuguese were searching for an alternate water route to India because

1. they wished to rediscover the route traveled by Marco Polo (4%).
2. the Turks had closed the old routes (59%).
3. the Spanish had proved that it was possible to reach the east by sailing westward (10%).
4. an all-water route would make possible greater profits (1% omitted the item) (26%).

More than half of the pupils selected Response 2. The negative index of discrimination indicates further that the average achievement of the pupils who selected this response was superior to that of the 25% of the pupils who selected the correct response (4). Authoritative historians no longer would accept the second response as a sufficient explanation of Portuguese attempts to round Africa, nor would they deny that Response 4 is the best of those given. An analysis of current textbooks in American history, however, will reveal that these lag behind research and that many of them still present the now-disproved explanation: "The Turks closed the old routes." It is not surprising, therefore, that the superior pupils are more likely to select this response than those who have made little or no effective attempt to learn the facts contained in the textbooks. This being the case, the inclusion of this item in the test not only contributed nothing to its effectiveness but also detracted from it. There can be little question, however, that the item is free from technical imperfections or ambiguities, and that it does hold the pupil responsible for an established fact of considerable significance in history.

Source: Adapted from The Construction and Use of Achievement Examinations: A Manual for Secondary School Teachers (pp. 56–63), by H. E. Hawkes, E. F. Lindquist, and C. R. Mann (Eds.), 1936, Boston: Houghton Mifflin, ©1936 by Houghton Mifflin Company. Adapted by permission of the publisher.

2. *Be sure that competent authorities can agree that the answer keyed as correct (or best) is in fact correct (or best).* If you violate this rule, you may come into conflict with the more able student (or the student's parent). Further, if you insist there is only one correct answer when students also see another choice as equally logical and correct, students will likely see you as arbitrary and capricious. To avoid such embarrassment and negative consequences, have a knowledgeable colleague review the correctness of your keyed answers and the incorrectness of your distractors before you use them. The best way to do this is to have your colleague take your test without the correct answers marked. If the colleague chooses an answer that you did not key as correct, then there may be a problem with the correctness of your key.

3. *The correct alternative should be a grammatically correct response to the stem.* The knowledgeable student faces a conflict if the content of the keyed response is correct, but the grammar is incorrect.

4. *Check over the entire test to ensure that the correct alternatives do not follow an easily learned pattern.* Use the answer key you develop to tabulate the number of A's, B's, C's, and so on that are keyed as correct. Sometimes teachers favor one or two positions (e.g., B and C) for the correct answers. Students will quickly catch on to this pattern, which lowers the validity of your assessment. Also, avoid repetitive, easily learned patterns, such as AABBCCDD or ABCDABCD.

5. *Avoid phrasing the correct alternative in a textbookish or stereotyped manner.* To assess comprehension and understanding, you must at least paraphrase textbook statements. Students quickly learn the idiosyncratic or stereotyped way in which you and the textbook phrase certain ideas. If your test items also reflect such idiosyncrasies, you will be encouraging students to select answers that "sound right" to them but that they do not necessarily understand. For more mature students, however, stereotyped phrases that have a "ring of truth" in the distractors may serve to distinguish those who have fully grasped the concept from those with only superficial knowledge (Ebel, 1972). Use this tactic with senior high school and college students, but not with elementary and junior high school students.

6. *The correct alternative should be of approximately the same overall length as the distractors.* Teachers sometimes make the correct option longer than the incorrect options by phrasing it in a more completely explained or more qualified manner. The testwise student can pick up on this and mark the longest or most complete answer without having the requisite knowledge. Research supports the generalization that if you violate this rule you will make the item easier (Haladyna & Downing, 1989b). Don't be too scrupulous in counting words, however. If your correct answer is one or two words longer, don't worry about it.

7. *An advantage of a multiple-choice test is that it reduces the amount of time required for writing answers, thus allowing the assessment to cover more content.* Don't defeat this purpose by requiring students to write out their answers. Have the students either mark (circle, check, etc.) the letter of the alternative they choose, write the letter on a blank next to the stem created for that purpose, or use a separate answer sheet. Separate answer sheets are not recommended for children below fourth or fifth grade. If your state has a testing program that uses separate answer sheets in the primary grades, however, use answer sheets with some of your classroom tests to give the children practice.

Encoding Meaning into Distractor Choices

Thus far we have discussed distractors that all serve the same purpose, namely, to appear plausible to those who do not know the correct answer. In scoring, all are equally "wrong." On a right-wrong, 1-0 item scoring scale, choosing a distractor gets a student 0 points. Several different programs of research have investigated encoding more meaning into distractors than simply "wrong."

It is possible to write distractors that help teachers identify what next steps a student should take. These can be based on cognitive developmental models of how children learn (Pellegrino, Chudowsky, & Glaser, 2001). So, for example, one of the distractors could represent what a student who is in the beginning stages of concept development would select, another distractor would represent what a student who has progressed to a second stage of concept development would select, and so on. In problem solving, distractors can be crafted to represent different kinds of mistakes. For example, for the problem 115−97 = ?, one of the distractors might be 22, which is what a student who always subtracted the smaller number from the larger might select. Another distractor might be 28, which is what a student who knew how to borrow in the one's place, but not how to change the value in the ten's place, would select.

Pearson Assessment has developed a **distractor rationale taxonomy** (King, Gardner, Zucker, & Jorgensen, 2004) for multiple-choice items in reading and mathematics. These taxonomies describe types of errors in reading and mathematics, respectively, that correspond with different levels of understanding. The advantage is that one distractor can be written for each level. A student whose incorrect answers are typically at a specific level can be given instruction targeted to that level of understanding. Figure 9.6 presents the distractor taxonomy for reading items and examples to illustrate its use. Readers who would like to see the mathematics distractor rationale taxonomy and additional examples should refer to the reference.

A Checklist for Evaluating Multiple-Choice Items

Practicing the preceding rules will help you write better multiple-choice items. It is difficult to keep all of the rules in mind, however. Some of the most useful rules are presented in the checklist. You can use this checklist to review the items you have written or those you have found in the quizzes and tests that come with your textbook or teaching materials. Revise every item that does not pass your checklist evaluation before you use it.

FIGURE 9.6 A distractor rationale taxonomy for reading items related to the main idea and vocabulary in context.

Level of understanding	Student error
LEVEL 1	Makes errors that reflect focus on decoding and retrieving facts or details that are not necessarily related to the text or item. Student invokes prior knowledge related to the general topic of the passage, but response is not text-based. These errors indicate that the student is grabbing bits and pieces of the text as he or she understands them, but the pieces are unrelated to the information required by the question being asked.
LEVEL 2	Makes errors that reflect initial understanding of facts or details in the text, but inability to relate them to each other or apply them to come to even a weak conclusion of inference. The student may be focusing on literal aspects of a text or on superficial connections to arrive at a response.
LEVEL 3	Makes errors that reflect analysis and interpretation, but conclusions or inferences arrived at are secondary or weaker than ones required for correct response. A distractor may be related to the correct response in meaning, but be too narrow or broad given the circumstances.
LEVEL 4	Correct response.

The examples are associated with a Grade 3 reading passage titled "Frogs and Toads."
The first example uses this taxonomy:

WHAT IS THE MAIN IDEA OF THE PASSAGE "FROGS AND TOADS"?

A. Frogs and toads are cute. [Level 1: prior knowledge, not text-based]
B. Toads have shorter legs than frogs have. [Level 2: text-based detail unrelated to main idea]
C. Frogs are different than toads. [Level 3: only part of main idea]
D. Frogs and toads share many differences and similarities. [Level 4: correct response]

The second example presents a traditional version of an item with the same stem, for contrast.

WHAT IS THE MAIN IDEA OF THE PASSAGE "FROGS AND TOADS"?

A. Frogs live closer to water than toads.
B. Frogs and toads are like cousins.
C. Frogs are different than toads.
*D. Frogs and toads share many differences and similarities.

All distractors are essentially Level 3: Each is related to the main idea but is not the best answer.

Source: The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics. Copyright © 2004 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

✔ **CHECKLIST**

A Checklist for Reviewing the Quality of Multiple-Choice Items

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does the item assess an important aspect of the unit's instructional targets?
2. Does the item match your assessment plan in terms of performance, emphasis, and number of points?
3. Does the stem ask a direct question or set a specific problem?
4. Is the item based on a paraphrase rather than words lifted directly from a textbook?
5. Are the vocabulary and sentence structure at a relatively low and nontechnical level?
6. Is each alternative plausible so that a student who lacks knowledge of the correct answer cannot view it as absurd or silly?
7. If possible, is every incorrect alternative based on a common student error or misconception?
8. Is the correct answer to this item independent of the correct answers of other items?
9. Are all of the alternatives homogeneous and appropriate to the content of the stem?
10. Did you avoid using "all of the above" or "none of the above" as much as possible?
11. Is there only one correct or best answer to the item?

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 35), by A. J. Nitko and T-C Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of copyright holders.

CREATING ALTERNATIVE VARIETIES OF MULTIPLE-CHOICE ITEMS

A number of multiple-choice item formats are usually not taught in traditional assessment courses, but they have considerable usefulness. The value of these item formats is fourfold. First, some of them will fit your learning targets much more closely than do typical true-false, matching, and multiple-choice formats, thus increasing the validity of your classroom assessments. Second, the formats are objectively scored. As you know, the more objective your scoring, the more likely you are to have reliable scores for evaluating your students. Third, because these tasks take students a relatively short time to complete, you can assess a wider range of content and learning targets by using one or more of these formats in addition to your traditional assessment formats. Fourth, these formats are relatively easy to create.

The section discusses four item formats: greater-less-same, best-answer, experiment-interpretation, and statement-and-comment. Many of the ideas for this discussion are adapted from Carlson (1985) and Gulliksen (1986). After each format is illustrated, we discuss advantages and criticisms, then offer suggestions for improving the way you craft the items.

Greater-Less-Same Items

The **greater-less-same item** format consists of a pair of concepts, phrases, quantities, and so on that have a greater-than, same-as, or less-than relationship. The greater-less-same item format is used to assess qualitative, quantitative, or temporal relationships between two concepts. Several examples are shown in Figure 9.7.

The student's task is to identify the relationship between the concepts and record an answer. You may use before-during-after, more-same-less, heavier-same-lighter, or other ordered triads, depending on the context of the items. Also, instead of spelling out the words *greater*, *less*, *same*, you can use the letters *G, L,* and *S,* respectively. Using letters instead of words may be more appropriate for older students.

Begin to create items by first identifying the learning targets you want to assess. This item format assesses learning targets that include the ability to identify the relationships between two ideas, concepts, or situations. You should make a list of concept pairs that are related; add to this list other paired relationships that your students can deduce from principles or criteria they have learned. Rephrase the members of each pair so they are clearly stated and fit the item format. When arranging the pairs, be sure that you do not have all the "greaters" on one side of the pair.

Write a set of directions for students that explains the basis on which they are to choose greater-same-less (before-during-after, etc.). Normally, *the set of items should refer to the same general topic*. In the examples in Figure 9.7 this is not the case, because we wanted to illustrate items from different subject areas. Therefore, the directions are too general. Your directions should be more focused on the set of items you are using and very clear. Notice, too, that Item 7 does not "fit" the directions.

The first time you use this format, you may need to give your students some sample items to

FIGURE 9.7 Examples of greater-less-same items.

Directions: The numbered items below contain pairs of statements. Compare the two members of each pair. If the thing described on the *left* is greater than the thing described on the right, circle the word "greater"; if the *left* is less than the right, circle "less"; and if the *left* and the right are essentially the same, circle "same."

1. Total area of Lake Erie	Greater Same [✓ Less]	Total area of Lake Huron
2. Meaning of the prefix *mono-*	Greater [✓ Same] Less	Meaning of the prefix *uni-*
3. Radius of Mars	Greater Same [✓ Less]	Radius of Venus
4. Number of Christians in Africa	[✓ Greater] Same Less	Number of Muslims in Africa
5. Atomic weight of Ca	[✓ Greater] Same Less	Atomic weight of C
6. $\sqrt{3^2 + 7^2}$	Greater Same [✓ Less]	$\sqrt{3^2} + \sqrt{7^2}$
7. First U.S. passenger railroad opened	Before Same [✓ After]	Erie Canal opened

help them understand what they are to do. Be sure the directions tell the students *which member* (i.e., *left* or *right member*) of the pairs in the set they are to use as a referent.

Organize all the items of this format into one section of your assessment. Put the directions and the sample item at the beginning of the set. The numbered items should follow. Be sure that the correct answers do not follow a set pattern (such as GSLGSL or GGLLSS). Review the set to be sure the items are concisely worded, the task is clear, and the relationships are not ambiguous.

The checklist that follows summarizes the suggestions in this section for judging the quality of greater-less-same items. Use the checklist to guide you in crafting this type of item format. Use it, too, to evaluate the item sets you have already crafted.

✔ CHECKLIST

A Checklist for Reviewing the Quality of Greater-Less-Same Items

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does each item in the greater-less-same set assess an important aspect of the unit's instructional targets?
2. Does each item in the greater-less-same set match your assessment plan in terms of performance, emphasis, and number of points?
3. Do some of the items in the greater-less-same set require students to apply their knowledge and skill to new situations, examples, or events?
4. Do your directions clearly and completely explain the basis you intend students to use when judging "greater than," "less than," or "same as" for each pair of statements?
5. Do your directions state which pair member (left or right) is the referent?
6. Did you avoid using a pattern (GGSSLLGGSSLL, etc.) for the correct answers?

Advantages The greater-less-same format is especially suited for assessing whether students understand the order or relationships between two concepts, events, or outcomes. These include greater than versus less than, more of versus less of, before versus after, more correct versus less correct, more preferred versus less preferred, heavier versus lighter, and higher quality versus lesser quality. When you teach the relationships in class or when students learn the relationships from the textbook, this item assesses recall and recognition. However, this item format need not be limited to recall or remembering. You may teach a principle or a set of criteria and give several examples of its application in class. Then, when assessing the students, *present new examples*. A student can then apply the principle(s) or criteria you taught to *deduce the relationship* between the concept pairs. This elevates the item so it requires a higher level of thinking than remembering.

Criticisms The criticisms of greater-less-same items are similar to those for matching and true-false items. That is, teachers often use them to assess rote association and disconnected bits of knowledge. Also, this format limits assessment to relationships among pairs of concepts. If you wish to assess a student's ability to order larger members of a set of events or facts, then use an item format that requires students to rank the members.

Best-Answer Items

Best-answer items are multiple-choice items for which every option is at least partly correct. The student's task is to select the best or most correct option. Here is an example of a best-answer item:

Example

Directions: The following question refers to the article below about the model United Nations General Assembly.

Text of article:

MODEL U.N. Coming to Town

Local students represent the United States (New York) Students from 15 countries from around the world will be arriving on Monday for a model session of the United Nations General Assembly. Each country will write a plan for the Assembly in one of four categories: Environment, Education, Culture, and Economic Development.

1. Which would be the best plan for the model United Nations General Assembly to improve the world environment?

   A Feature different ethnic foods in the cafeteria next week.
   B Plan a school lunch program in Chicago.
   C Give food to refugees from war zones.
   *D Study the effect of acid rain on crops.

Source: National Assessment of Educational Progress, released item: Civics, grade 4, block 2006-4C3, no. 9. Available: http://nces.ed.gov/nationsreportcard/itmrls/

In this item format, each distractor contains partial misinterpretations or omissions. The keyed or best answer contains neither misinterpretations nor omissions. Only one option can be the "best." Therefore, you should never use "all of the above" or "none of the above" with this format. Neither can some combination of choices (such as "both A and C") be the keyed answer.

As always, first identify the learning targets you want to assess. Learning targets that require students to choose among several partially correct alternatives may be assessed using this format. Before using this type of item, be sure you have taught your students to use criteria for selecting the best among several partially correct explanations, descriptions, and ideas. These are higher-order thinking skills (often called critical-thinking skills) in that students must use criteria (such as "completeness of response" and "no misinformation") to evaluate alternatives.

Begin by first drafting the question for the stem. Second, write several ways in which students' responses to that question are typically partially correct. These become the basis for writing distractors. You could also give your students several open-ended short-answer questions as homework. Then, select from among the students' responses those that represent excellent, good, and poor answers. Edited versions of these could be used as a basis for creating the options. (Do not use students' responses verbatim as alternatives. They may be poorly phrased or contain too many other errors to function well as partially correct distractors.)

Because the best-answer format is a multiple-choice format, you should follow the basic rules in the checklist for evaluating the quality of multiple-choice items. A typical flaw with best-answer items is that the best or keyed answer is the one with the longest wording because it contains the most complete information. Avoid this flaw by being sure the options have approximately equal numbers of words.

Use the following checklist for judging the quality of best-answer items. Use it, too, as an evaluation guide as you review and edit the items you have already created.

✔ CHECKLIST

A Checklist for Reviewing the Quality of Best-Answer Items

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does each best-answer item assess an important aspect of the unit's instructional targets?
2. Does each best-answer item match your assessment plan in terms of performance, emphasis, and number of points?
3. Does each best-answer item require students to apply their knowledge and skill in some manner to new situations, examples, or events?
4. Do your directions clearly and completely explain the basis you intend students to use when judging "best"? (Have your students been given practice in using the appropriate criteria for judging "best"?)
5. Are all the options correct to some degree?
6. Is the keyed answer the only one that can be defended as "the best" by applying the criteria you specify in the directions?
7. Is each distractor based on an important misconception, misunderstanding, or way of being an incomplete answer? (Did you avoid tricky or trivial ways of making a distractor partially correct or contain misinformation?)
8. Are all of the options of equal length (within five words of each other)?
9. Did you avoid (a) having more than one "best" answer and (b) using "all of the above" or "none of the above"?
10. Did you apply all of the multiple-choice item-writing guidelines described in the multiple-choice checklist?

Advantages Best-answer items assess students' ability to make relatively fine distinctions among the choices. They must comprehend the question and the criteria used to judge the "best" option. Thus, best-answer items assess relatively high-order verbal reasoning skills.

Criticisms Best-answer items are difficult to write. You must know your subject and your students' faulty thinking patterns quite well. You need to create distractors that are partially correct, yet less defensible than the keyed answers. This is unlike typical multiple-choice items for which one option is the only correct one and the others are incorrect. Another criticism is that this format may be unsuitable for some students because their level of educational development is not high enough to make the fine distinctions necessary to select the best answer.

A third criticism is that different teachers may not teach consistently across sections of the same course. Thus, what is legitimately a best answer in one teacher's class might not be the best answer in another teacher's class. A fourth criticism is that "best" implies a set of criteria that students may not have been taught or may fail to understand. No answer is unequivocally best unless it is evaluated by applying these criteria. Your students must internalize criteria to apply them. Also, your own knowledge of the subject may be limited. As a result, what you consider the best answer may in fact not be best, because you do not understand other criteria by which the options may be evaluated. A fifth criticism is that a teacher may easily write a tricky item—that is, an item in which an option's correctness depends on a trivial fact, an idiosyncratic standard, or an easily overlooked word or phrase.

Experiment-Interpretation Items

The **experiment-interpretation item** consists of a description of an experiment followed by a multiple-choice item requiring students to recognize the best interpretation of the results from the experiment. Below are three examples. Items 1 and 2 are for a unit in general or physical science; Items 3–6 are for a social studies unit or a mathematics unit on statistical methods. We use the term *experiment* loosely in this section to mean any data-based research study. Scientific or controlled studies are included in the term, but we do not limit its use to only those types of studies. The experiment-interpretation item is similar to the best-answer format because very often the multiple-choice options will all have some degree of correctness, but only one is the best answer. A variation is to use a short-answer item along with or instead of the multiple-choice items (see Example Items 3–6). For example, you may ask a student to justify her choice on the multiple-choice item. Alternatively, you could use a short-answer question instead of the multiple-choice one.

Examples

Use the following information to answer Question 1.

Billy and Jesse were walking through an empty lot near their home. Billy picked up a whitish rock.

"Look," he said, "I found a limestone rock. I know it is a limestone rock because I found a rock last year that has the same color and it was limestone."

Jesse said, "Just because it looks the same it doesn't have to be the same."

1. Which of the following explanations best supports *Jesse's* point of view?
   A During the year the chemical properties of limestone probably changed.
   B Different minerals have very similar physical properties.
   C One year is not long enough for the minerals in a rock to change their physical properties.

Use the following information to answer Question 2.

Billy took the rock home and did an experiment with it. He put a piece of the rock in a clear glass and poured vinegar over it. The piece of rock bubbled and foamed. "There!" he said to Jesse, "That proves the rock is limestone."

Jesse said, "No! You are wrong. You haven't proved it!"

2. Why was Jesse correct?
   A Billy did the experiment only once. He needs to repeat the same type of experiment many times with different bits of the rock. If the mixture bubbles every time, that will prove it.
   B The experiment is correct but Billy misinterpreted the results. Limestone does not bubble and foam in vinegar.
   C Billy should do many different kinds of experiments, not just vinegar tests, because many different kinds of substances bubble and foam in vinegar.
   D Billy should not have used vinegar. He should have used distilled water. If the rock made the water warm, that would prove it is limestone.

Use the following information when answering Questions 3 to 6.

For a social studies project, a class interviewed all the 10th-grade students. They asked how many hours per week students worked at after-school jobs. They also asked what their average grades were last term. They found that students with Fs and Ds worked 8 to 10 hours per week, students with Cs and Bs worked 10 to 20 hours per week, and students with As worked 8 to 10 hours per week.

Alternative Format A

Students choose from among teacher-provided interpretations but are required to write a justification of their choice.

3. Which of the following is the most valid interpretation of these findings?
   A If you work 10 to 20 hours per week you will only get Cs and Bs.
   B Working after school is not related to your grades.
   *C A student who works 10 to 20 hours per week is probably not an A student.
   D The more hours a student works after school, the higher will be that student's grades.

4. Write a brief explanation of why your answer to Question 3 is the most valid interpretation of these findings.

____

____

____

____

Alternative Format B

Students supply their own interpretation and justify it in writing.

5. What is the most valid interpretation of the relationship the class found between the number of hours students worked and their grades?

____

____

____

____

6. Write a brief explanation of why your interpretation of these findings is the most valid one.

____

____

____

____

The three variations (multiple-choice only, multiple-choice with short-answer, and short-answer only) in the social studies/mathematics example assess somewhat different abilities. Using multiple-choice only (Item 3) assesses a student's ability to evaluate each of *the interpretations you provide* and select the best one. Thus you do not know a student's reasoning behind his selection. The multiple-choice with short-answer combination (Items 3 and 4) assesses a student's ability to explain or justify her choice from among the interpretations you provide as options. This helps you assess the reasoning behind students' choices. The short-answer *without the multiple-choice items* (Items 5 and 6) assess both a student's ability to interpret the experiment's results and his ability to explain his reasoning. In this latter format, there may be multiple correct responses to the constructed-response questions. As with other constructed-response items, you may want to give students partial credit if their response is not completely correct.

First, identify the learning targets you want to assess. The experiment-interpretation assessment format is appropriate when a learning target requires students to understand and interpret the results of empirical research. Before using this format for summative student evaluation, be sure you have taught and have given practice in interpreting the findings from empirical research studies.

Write the item to assess the student's ability to apply specific principles. This means that you first identify the principles or rules you want students to apply, then craft the item so it requires students to use the principle in a new situation. For example, items in the preceding examples are crafted around the following principles:

- Different substances may share the same or similar physical properties such as color, texture, and solubility. [Item 1]
- Different substances may share the same or similar chemical properties, such as their reactivity with acids. [Item 2]
- Some patterns of relationships among variables are not strictly increasing or decreasing but are curvilinear. [Items 3 through 6]

After identifying the principle(s), you create the item in such a way that it requires students to use or apply the principle(s). Usually, this means writing a description of the experiment or research study that results in findings that a student can then interpret using the principle(s). (See the interpretive text that immediately precedes Items 1, 2, 3/4, and 5/6 in the previous examples.)

Next, draft a stem that asks the student to interpret or explain the experimental findings you describe. You may then list several correct or partially correct interpretations. You may also list incorrect interpretations that result from incomplete or faulty reasoning. Avoid using as distractors interpretations that are completely unrelated to the experiment you describe in the interpretive material or distractors that are "silly" or "tricky." For example, it would be inappropriate for you to use in Item 1 a distractor such as "Jesse knows that Billy is a liar."

As with the best-answer item format, distractors for this format should contain interpretations or explanations that contain your students' typical misconceptions. To determine these misinterpretations, you could assign several open-ended questions as homework and select from among the students' responses those that are excellent, good, and poor. Use these selections as a basis for creating multiple-choice options.

If you use the multiple-choice versions of this format, you should follow the basic rules of writing multiple-choice items that we discussed earlier and that are summarized in the multiple-choice checklist. If you use one of the short-answer versions of this format, you should follow the basic rules of short-answer item writing (Chapter 8). The following checklist offers specific guidance for the experiment-interpretation item format. Use it to review the items you craft.

✔ **CHECKLIST**

A Checklist for Reviewing the Quality of Experiment-Interpretation Items

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does each item assess an important aspect of the unit's instructional targets?
2. Does each experiment-interpretation item match your assessment plan in terms of performance, emphasis, and number of points?
3. Does each item focus on requiring students to apply one or more important principles or criteria to new situations, examples, or events?
4. Have you given students opportunity to practice applying the appropriate criteria or principles for judging the "best" or "most valid" interpretation?
5. Did you describe an experiment or research study in concise but sufficient detail that a student can use the appropriate criteria or principles to interpret the results?
6. Is the keyed answer the only one that can be defended as the "best" or "most valid" interpretation?
7. Is each distractor based on an important misconception, misinterpretation, or misapplication of a criterion or principle? Did you avoid tricky or trivial ways of making a distractor partially correct or contain misinformation?
8. Did you avoid (a) having more than one "best" or "most valid" answer and (b) using "all of the above" or "none of the above"?
9. Did you apply all of the appropriate item-writing guidelines described in the multiple-choice checklist?

10. If you used short-answer items, did you apply all of the appropriate item-writing guidelines described in the short-answer checklist?

Advantages You may use the experiment-interpretation format to assess a student's ability to evaluate explanations, interpretations, and inferences from data. The multiple-choice-only version allows you to score the items more quickly and more objectively than the other versions. Because students are required only to select the correct answer, their response times are shorter. Therefore, you can use more items and cover more content within a shorter assessment period than with short-answer items.

If the experiments and findings you present in the items are new to the students, your items will assess your students' ability to apply principles and criteria from your subject area. Using experiments and data new to your students in assessment tasks requires you to teach students how to apply criteria and principles to a variety of situations. You will need to give students sufficient practice in applying criteria and principles before assessing them for summative evaluation purposes. This will move your teaching away from teaching facts and results, and toward teaching students to actively apply their knowledge and skill.

If you require students to justify their multiple-choice answers, you will have some information about their reasoning processes. Students often make the correct choice from among the possible interpretations you give them, but they cannot explain why they made the choice, or they give faulty explanations. If you require a student both to supply his interpretation and to justify it, you can assess whether the student can generate and explain his own interpretations of experimental findings.

Criticisms Like the best-answer item format, the experiment-interpretation format is not easy to write. You must know your subject matter and your students' thinking patterns well enough to create items that allow you to identify faulty thinking as well as correct answers. Faulty thinking must be reflected in your multiple-choice distractors. This means you must be able to create partially correct interpretations and incorrect interpretations that people typically make.

Use experiment-interpretation items to assess higher-order thinking. Do not use this format to assess whether a student can remember the "correct" interpretations of specific experimental results you taught. Using this format to assess remembering encourages students to look to the teacher or the text as the source of fixed knowledge. It discourages students from learning skills required to interpret the empirical results of experiments.

Statement-and-Comment Items

A **statement-and-comment item** presents a statement about some relevant subject matter and requires the student either to write a comment about the statement or to select the most appropriate comment from among a list you provide. Here is an example of a statement-and-comment item:

Example

A. Multiple-choice version of a statement-and-comment item

The Bundle of Sticks—Aesop An old man on the point of death summoned his sons around him to give them some parting advice. He ordered his servants to bring in a bundle of sticks, and said to his eldest son: "Break it." The son strained and strained, but with all his efforts was unable to break the bundle. The other sons also tried, but none of them was successful. "Untie the bundle," said the father, "and each of you take a stick." When they had done so, he called out to them: "Now, break," and each stick was easily broken. "You see my meaning," said their father.

Directions: The quote expresses the theme of Aesop's fable "The Bundle of Sticks." Choose the answer that best expresses how the theme applies to the fable.

1. "Union gives strength."
   A The three sons all tried to break the bundle.
   *B None of the sons could break the bundle of sticks.
   C Each of the sons could break a single stick.

B. Short-answer version of a statement-and-comment item

Directions: The quote expresses the theme of Aesop's fable "The Bundle of Sticks." Below the quote, explain how the theme applies to the fable.

2. "Union gives strength."

____

____

____

In the multiple-choice version, a student selects from among several alternate choices the best meaning of the quoted theme. The multiple-choice version is a special case of the best-answer item format. The alternatives should be phrased in language different from the "pat phrases" learned in class. In the short-answer version, students must comment directly, writing their own interpretation of the quoted statement.

First, as always, identify the learning targets you want to assess. This assessment format is appropriate when a learning target requires a student to comprehend statements and themes.

If you give students the short-answer version as a homework exercise, you may use excellent, good, and poor student responses as a basis for creating the alternatives for the multiple-choice version. As with the best-answer variety, of which this may be considered a special case, you usually cannot use students' responses verbatim as multiple-choice options; paraphrase them. Because the multiple-choice version of the statement-and-comment is a type of best-answer item, follow the guidelines suggested in the best-answer item checklist.

Advantages The statement-and-comment item format assesses a student's ability to evaluate interpretations of a given statement. The multiple-choice version assesses whether students can identify the best interpretation from among several. Interpretations should not use the same wording used in class. Rather, they should be comments typically made by students when interpreting the quoted statement. In this way, students must rely on their comprehension of the quoted phrase instead of their memory of a "set" comment.

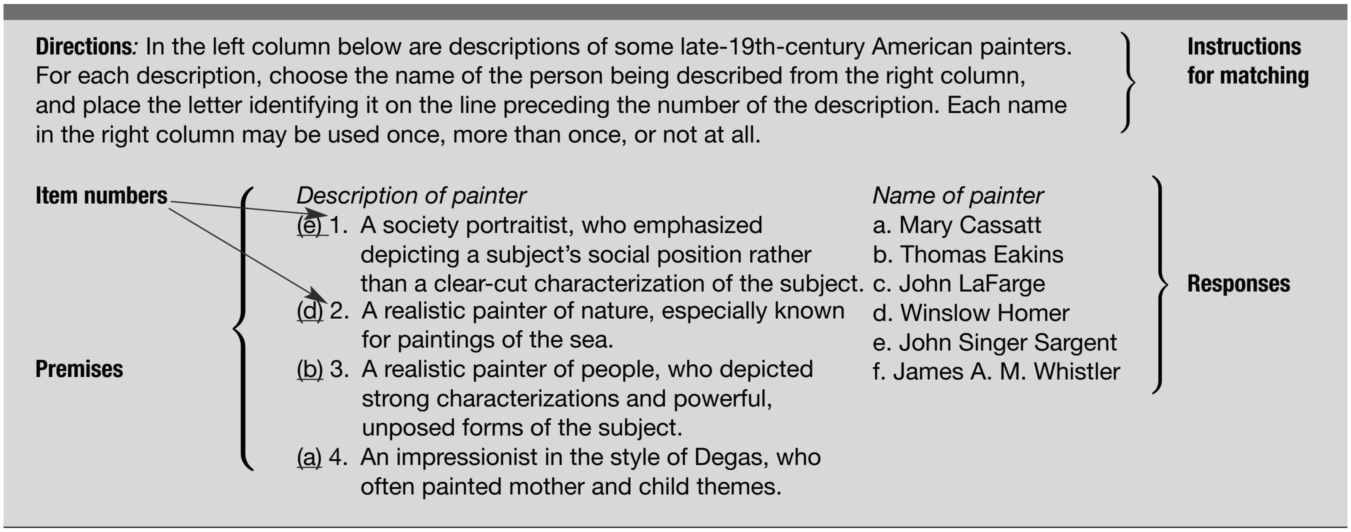
The open-ended version assesses students' ability to recall and write about the meaning of the quoted statement. Although it may be an advantage to have students construct their own comments about the quoted statement, there is a downside. Students may just write an explanation or commentary they memorized from the class discussion or from a textbook. You have some control over what kinds of comments they must evaluate if you present the multiple-choice version.

Criticisms The statement-and-comment item format has limited applications. You must identify appropriate statements that students should interpret. Although there are many subjects for which such statements exist, the task itself represents a small range of learning targets. The short-answer version of the task does provide an opportunity for students to display their comprehension of the quoted statement. However, students may simply repeat the phrases they learned in class.

MATCHING EXERCISE FORMAT

A **matching exercise** presents a student with three things: (1) **directions for matching,** (2) a **premise list**, and (3) a **response list**. The student's task is to match each premise with one of the responses, using as a basis for matching the criteria described in the directions. Figure 9.8 shows a matching exercise with its various parts labeled.

FIGURE 9.8 **Example of a matching exercise.**



Directions*:* In the left column below are descriptions of some late-19th-century American painters. For each description, choose the name of the person being described from the right column, and place the letter identifying it on the line preceding the number of the description. Each name in the right column may be used once, more than once, or not at all.

Description of painter

(e) 1. A society portraitist, who emphasized depicting a subject's social position rather than a clear-cut characterization of the subject.

(d) 2. A realistic painter of nature, especially known for paintings of the sea.

(b) 3. A realistic painter of people, who depicted strong characterizations and powerful, unposed forms of the subject.

(a) 4. An impressionist in the style of Degas, who often painted mother and child themes.

Name of painter

a. Mary Cassatt
b. Thomas Eakins
c. John LaFarge
d. Winslow Homer
e. John Singer Sargent
f. James A. M. Whistler

The sample exercise in Figure 9.8 requires simple matching based on associations that students must remember. You may create matching exercises, however, to assess students' comprehension of concepts and principles. Examples of these latter types appear later in the chapter.

In matching exercises, premises are listed in the left column and responses in the right column, or responses are listed vertically above the premises. Each premise is numbered because each is a separately scorable item. Matching exercises can have more responses than premises, more premises than responses, or an equal number of each. **Perfect matching** occurs when you have an equal number of premise statements and response statements. Most assessment specialists consider perfect matching to be undesirable because, if a student knows four of the five answers, the student automatically gets the fifth (last) choice correct, whether or not he knows the answer. This reduces the validity of the assessment results.

Matching exercises are very much like multiple-choice items. Each premise functions as a separate item. The elements in the list of responses function as alternatives. You could rewrite a matching exercise as a series of multiple-choice items: Each premise would then be a multiple-choice stem, but the same alternatives would be repeated for each of these stems. This leads to an important principle for crafting matching exercises: *Use matching exercises only when you have several multiple-choice items that require repeating the identical set of alternatives.*

ADVANTAGES AND CRITICISMS OF MATCHING EXERCISES

Advantages

A matching exercise can be a space-saving and objective way to assess a number of important learning targets, such as your students' ability to identify associations or relationships between two sets of things. You can also develop matching exercises using pictorial materials to assess the students' abilities to match words and phrases with pictures of objects or with locations on maps and diagrams. Figure 9.9 gives examples of relationships that you may use as a basis for developing matching exercises.

FIGURE 9.9 Examples of different foundations for developing matching exercises.

Possible premise sets	Associated response sets
Accomplishments	Persons
Noted events	Dates
Definitions	Terms and phrases
Examples, applications	Rules, principles, and classifications
Concepts (ideas, operations, quantities, and qualities)	Symbols and signs
Titles of works	Authors and artists
Foreign words and phrases	English correspondence
Uses and functions	Parts and machines
Names of objects	Pictures of objects

Criticisms

Detractors criticize the matching exercise because students can use rote memorization to learn the elements in two lists, and because teachers often use matching exercises only to assess such rote associations as names and dates. As a result, critics often see this assessment format as limited to the assessment of memorized factual information.

Thoughtful teachers, however, also use matching exercises to assess aspects of students' comprehension of concepts, principles, or schemes for classifying objects, ideas, or events (we will see examples later). *If you want to assess students on these higher-level abilities, create exercises that present new examples or instances of the concept or principle to the students.* Then require students to match these examples with the names of appropriate concepts or principles. In this context, *new examples* are instances of concepts that students have not been previously taught or encountered. Similarly, a matching task can describe a situation novel to the student, and the student can decide which of several rules, principles, or classifications is likely to apply. An example of this type of matching exercise follows:

Example

Directions: Each numbered statement below describes a testing situation in which ONE decision is represented. On the blank next to each statement, write the letter:

A if the decision is primarily concerned with placement
B if the decision is primarily concerned with selection
C if the decision is primarily concerned with program improvement
D if the decision is primarily concerned with theory development
E if the decision is primarily concerned with motivating students

(A)1. After children are admitted to kindergarten, they are given a screening test to determine which children should be given special training in perceptual skills.

(A)2. At the end of the third grade, all students are given an extensive battery of reading tests, and reading profiles are developed for each child. On the basis of these profiles, some children are given a special reading program, whereas others continue on with the regular program.

(B)3. High school seniors take a national scholastic aptitude test and send their scores to colleges they wish to attend. On the basis of these scores, colleges admit some students and do not admit others.

(E)4. Students are informed about the learning targets their examination will cover and about how many points each examination question will be worth.

This exercise assesses a student's understanding of five concepts related to using tests for decision making. The placement of the response list above the premise list creates a type of matching exercise called the **masterlist variety**. It is also called the **classification** or **keylist variety**. Later in this chapter we present suggestions for creating this type of matching exercise as well as the double-matching exercise or tabular exercise.

Using **homogeneous premises and responses** means that the elements in the premise list and the elements in the response list together refer to the same category of things. In the preceding example, for instance, all premises and responses refer to some type of educational decision. In the example in Figure 9.8, all premises and responses refer to late-19th-century painters.

Why should you create matching exercises with homogeneous premises and responses? Because the *entire* list of response choices has to be plausible for *every* premise. If it is not, the students' matching task may be trivial. As an example, consider the nonhomogeneous, poor-quality matching exercise shown here:

Example

Poor: Premises and response set are not homogeneous

(d) 1. Pennsylvania's official state flower	a. Ruffed grouse
(a) 2. Pennsylvania's official state bird	b. Pittsburgh
(b) 3. Major steel-producing city in the 1940s	c. 1,950,098
(c) 4. 1970 population of Philadelphia	d. Mountain laurel
	e. Allegheny River

Not all of the responses in this example are plausible distractors for *each* premise. As a result, students can answer the items on the basis of general knowledge of a few of the associations and common sense, rather than on any special knowledge learned from the curriculum.

This matching exercise is poor for another, perhaps more important, reason: *The main focus of the exercise seems lost.* Even if you tried to improve it, your efforts would probably be self-defeating. You may attempt to make the exercise's responses more homogeneous, but this may result in an exercise that does not assess the intended learning target. For example, you could make all the premises refer to different states and all the premises to different official state birds: The task would be to match the birds with the states. Your local curriculum, however, may only require students to identify their own state bird (or other facts and symbols about their own state). Creating a homogeneous exercise, as in this example, may result in a test that does not match the curriculum and, therefore, cannot be used. Remember that the learning targets determine the type of assessment.

Could anything be done to salvage this exercise? Remember the rule mentioned earlier in this chapter: You should reserve the matching exercise for situations when several multiple-choice items require the same set of responses. Returning to the example, note that each premise could be turned into a separate multiple-choice item, each with a different set of plausible options. Plausible options for a multiple-choice item on Pennsylvania's official state flower, for example, would include flowers native to the Pennsylvania region (e.g., daisies, roses, violets, etc.). Similarly, separate multiple-choice items could assess knowledge of the official state bird, names of cities, and size of cities.

CREATING BASIC MATCHING EXERCISES

Many of the suggestions for writing multiple-choice items apply to matching exercises as well. A few maxims, however, apply particularly to matching exercises. If you follow these, your assessment quality will improve. These suggestions are summarized in a checklist and discussed here. You should use the checklist to evaluate your own matching exercises or those that you adapt from teachers' texts or other curricular materials.

✔ **CHECKLIST**

A Checklist for Reviewing the Quality of Matching Exercises

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does the exercise assess an important aspect of the unit's instructional targets?
2. Does the exercise match your assessment plan in terms of performance, emphasis, and number of points?
3. Within this exercise, does every premise and response belong to the same category of things?
4. Do your directions clearly state the basis you intend students to use to complete the matching correctly?
5. Does every element in the response list function as a plausible alternative to every element in the premise list?
6. Are there fewer than 10 responses in this matching exercise?
7. Did you avoid "perfect matching"?
8. Are the longer statements in the premise list and the shorter statements (names, words, symbols, etc.) in the response list?
9. If possible, are the elements in the response list ordered in a meaningful way (logically, numerically, alphabetically, etc.)?
10. Are the premises numbered and the responses lettered?

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 34), by A. J. Nitko and T.-C. Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of copyright holders.

Crafting Suggestions

1 and 2. *As always, your assessment tasks should meet the dual criteria of importance and fit with your assessment plan.* Eliminate every item that fails to meet these two criteria.

3. *Create homogeneous matching exercises.* We have discussed this point previously. Remember that the degree to which students perceive the exercise as homogeneous varies with their maturity and educational development. What may be a homogeneous exercise for primary schoolchildren may be less so for middle school youngsters and even less so for high schoolers. Consider, for example, the following matching exercise:

Example

Directions: Column A below lists important events in U.S. history. For each event, find in Column B the date it happened. Write the letter of the date on the blank to the left of each event. Each date in Column B may be used once, more than once, or not at all.

Column A (events)	*Column B (dates)*
(f) 1. United States entered World War I	a. 1492
(d) 2. Lincoln became president	b. 1607
(g) 3. Truman became president	c. 1776
(b) 4. Pilgrims landed at Cape Cod	d. 1861
	e. 1880
	f. 1917
	g. 1945

The students' task is to match U.S. historical events with their dates. For younger, less experienced students, such a matching task would likely be difficult. It would appear homogeneous, however, because for these children all responses would be plausible options for each premise. High school students would find the task easier—even though they didn't know the exact dates—because they could use partial knowledge to organize the dates into early, middle, and recent history. For them, only Options f and g would be plausible for Item 1.

4. *Explain completely the intended basis for matching.* You must make clear what basis you want students to use to match the premises and the responses. The example below shows how to improve the directions by explaining the basis for the matching:

Example

Poor: Directions are Incomplete

Match Column A with Column B. Write your answer on the blank to the left.

Better: Directions explain basis for matching

Column A lists parts of a plant cell. For each cell part, choose from Column B the main purpose of that cell part. Write the letter of that purpose on the blank to the left of the cell part.

Elementary students may need oral explanations and, perhaps, some practice with this format before you assess them. The masterlist variety of matching exercise usually requires more elaborate directions and may require special oral explanations even for high school students. Avoid long, involved written directions, however. These place an unnecessary premium on reading skill.

5. *All responses should function as plausible options for each premise.* Homogeneous premises and responses will minimize plausibility problems. Also, avoid using specific determiners and grammatical clues. For example, avoid beginning some premises or responses with *an* and others with *a,* having some plural whereas others are singular, stating some in the past tense whereas others are stated in the present or future tense. These clue the answer unnecessarily.

Avoid using incomplete sentences as premises. This makes it difficult to make all responses homogeneous and easier for students to respond correctly on the basis of superficial features such as grammatical clues or sentence structure. Here is an example of a poor matching exercise that comes about when incomplete sentences are used:

Example

Poor: Uses incomplete sentences

(c) 1. Most normally green plants lose their color when

(e) 2. The common characteristic of a flowering plant is

(d) 3. Almost all plants that form coal

(b) 4. When an expanded amoebae is strongly stimulated it

a. through their stomata.
b. contracts into a rounded mass.
c. grown in the dark.
d. are now extinct.
e. the formation of a reproductive body.

Source: From *The Construction and Use of Achievement Examinations* (p. 69), by H. E. Hawkes, E. F. Lindquist, and C. R. Mann (Eds.), 1936, by American Council of Education. Used by permission. Adapted from an illustration in *Traditional Examinations and New Type Tests* (p. 380), by C. W. Odell, 1928, New York: Century Co.

6. *Use short lists of responses and premises.* For a single matching exercise, put no more than 5 to 10 elements in a response list. The reasons are that (a) longer lists make it difficult for you to develop homogeneous exercises, (b) longer matching exercises overload a test with one kind of performance, (c) longer lists require too much student searching time, and (d) students may attain a lower percentage of correct answers with longer matching exercises than with shorter exercises.

Shorter matching exercises make it easier to keep everything belonging to a single exercise on the same page. For some students, having to turn the page back and forth to answer the exercise may interfere with their ability to show you what they know. For these students, splitting an exercise between two pages increases the likelihood of carelessness, confusion, and short-term memory lapses. In short, a student's ability to answer a test item while flipping pages is not relevant to the learning target you want to assess.

To fix an exercise that is too long, you can separate it into two or more shorter exercises. Or you can use each response as a correct answer more than once. When you do this, alert students through either oral or written directions. One standard phrase you may use to do this is "You may use each of the [names, dates, etc.] once, more than once, or not at all" (see the painters' example in Figure 9.8 at the beginning of this section).

7. *Avoid "perfect matching."* As we discussed previously, perfect matching is undesirable. It gives away at least one answer to the student who knows all but one of them. This student's final choice will be automatically correct because it is the only one left, thus lowering the validity of your assessment. You can avoid perfect matching by including one or more responses that do not match any of the premises and by using a response as the correct answer for more than one premise.

8. *Use longer phrases in the premise list, shorter phrases in the response list.* Consider how a student approaches the matching exercise: (a) first reading a premise, (b) then searching through the response list for the correct answer, and (c) rereading the response list for each premise. It is, therefore, more efficient and less time-consuming if students read the longer phrases only once. They can reread or scan the shorter phrases (words, symbols) as often as necessary.

9. *Arrange the response list in a logical order.* A student saves time if the response list is arranged in some meaningful order: Dates arranged chronologically, numbers in order of magnitude, words and names alphabetically, and qualitative phrases in a logical sequence. Such arrangements also may

contribute to the clarity of the task, reduce student confusion, and lower incidence of student carelessness and oversight.

10. *Identify premises with numbers and responses with letters.* Remember, each premise is a separately scored item. Therefore, premises should carry numbers, which indicate their position in the sequence of items. For example, if the first 10 items are multiple-choice, and these are followed by a five-premise matching exercise, the five premises should be consecutively numbered 11 through 15.

CREATING ALTERNATIVE VARIETIES OF MATCHING EXERCISES

Two types of matching exercises—masterlist and tabular—may fit some of your learning targets better than the more basic matching exercise. As with the alternative varieties of multiple-choice items we discussed previously, these matching formats are objectively scored, do not take students a long time to complete, and are often easier for you to craft than the basic matching format.

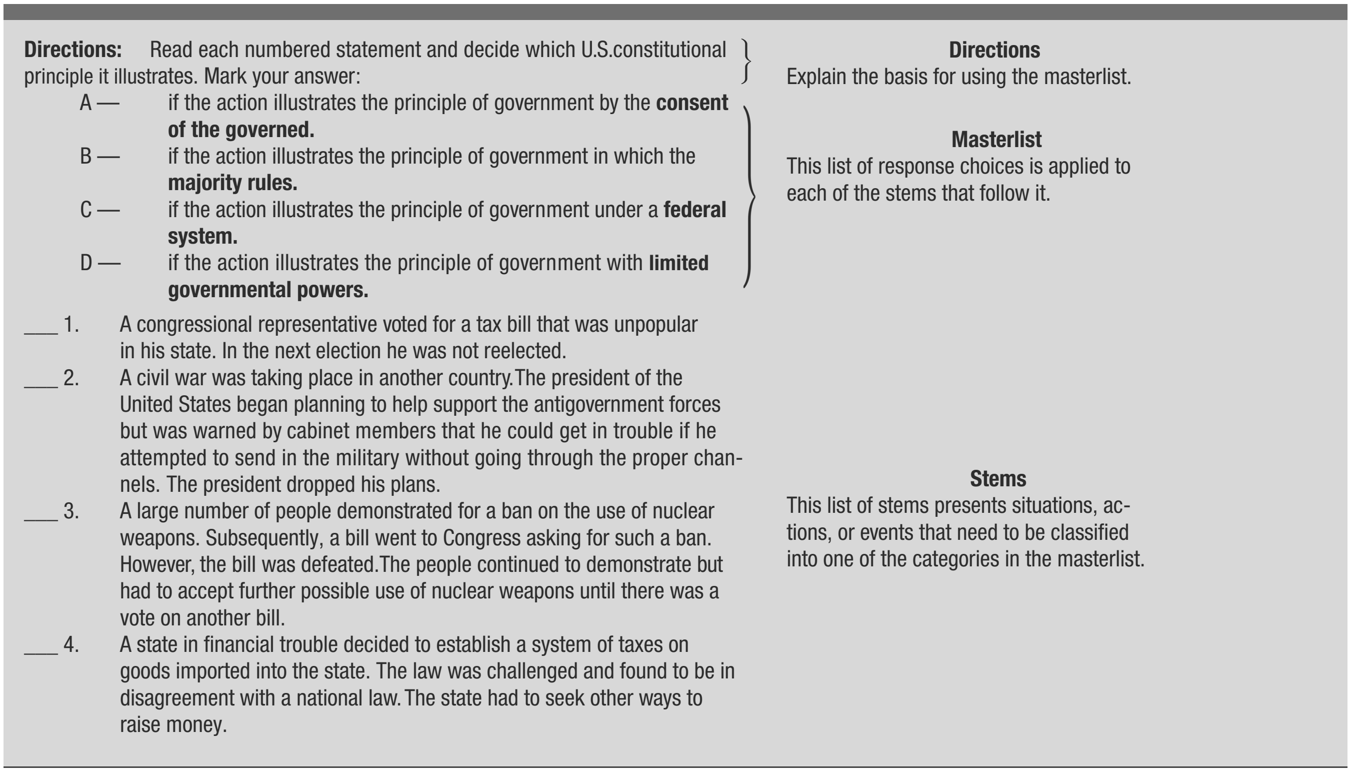
Masterlist (Keylist) Items

A masterlist (or keylist or classification) matching exercise has three parts: (1) directions to students, (2) the masterlist of options, and (3) a list or set of stems. To respond to a masterlist item set, a student reads each numbered stem and applies one of the options from the masterlist. Each stem is scored separately. Figure 9.10 shows a masterlist matching exercise for a 10th-grade civics course.

The content learning target for this masterlist exercise is the students' ability to relate constitutional *values and principles* to specific modern-day examples of actions or events. Therefore, in crafting this item you would ensure that each masterlist response choice (A, B, C, D) is a value or principle expressed by the U.S. Constitution, rather than a Preamble goal or some other aspect of the Constitution.

Notice that each numbered stem is a brief, realistic, and concrete example of an action or event that illustrates one of the four values in the masterlist. Because the learning target calls for students to relate constitutional principles to concrete examples, each stem must be a concrete example. You would not use textbook abstractions or general descriptions. (For example, you would not word a stem in general language such as, "A law takes effect when the majority of Congress votes to approve it," because this statement describes a general principle rather than a concrete example.)

FIGURE 9.10 **Example of a masterlist matching exercise.**



Directions: Read each numbered statement and decide which U.S.constitutional principle it illustrates. Mark your answer:

A — if the action illustrates the principle of government by the **consent of the governed.**

B — if the action illustrates the principle of government in which the **majority rules.**

C — if the action illustrates the principle of government under a **federal system.**

D — if the action illustrates the principle of government with **limited governmental powers.**

___ 1. A congressional representative voted for a tax bill that was unpopular in his state. In the next election he was not reelected.

___ 2. A civil war was taking place in another country.The president of the United States began planning to help support the antigovernment forces but was warned by cabinet members that he could get in trouble if he attempted to send in the military without going through the proper channels. The president dropped his plans.

___ 3. A large number of people demonstrated for a ban on the use of nuclear weapons. Subsequently, a bill went to Congress asking for such a ban. However, the bill was defeated.The people continued to demonstrate but had to accept further possible use of nuclear weapons until there was a vote on another bill.

___ 4. A state in financial trouble decided to establish a system of taxes on goods imported into the state. The law was challenged and found to be in disagreement with a national law.The state had to seek other ways to raise money.

Directions
Explain the basis for using the masterlist.

Masterlist
This list of response choices is applied to each of the stems that follow it.

Stems
This list of stems presents situations, actions, or events that need to be classified into one of the categories in the masterlist.

Although the preceding exercise shows only four stems, you need not limit the stems to four. Use as many stems as are appropriate, as long as each stem is an example of one of the masterlist options. Further, although not the case in this example, each stem in a masterlist set may have more than one correct answer from the masterlist.

To create your masterlist item set, first identify the learning target you want to assess. For example, this might be "the students' ability to recognize whether data support interpretations about what events occurred." Next write the masterlist of options on which you want to focus. For example, for the constitutional principles exercise given previously, you would list the four constitutional principles; for the masterlist exercise in Figure 9.11, you would list *supportive, contradictory,* and *neither*. If you will use a table, graph, or other interpretive material, prepare it next.

Select one of the options from the masterlist and write as many stems for it as you can. For example, you might select "consent of the governed" as a principle and write four or five concrete examples that illustrate that principle in a real-world application. Continue selecting options and writing stems until you have several items for each option. Review the stems to be sure that they require students to apply their knowledge and skills to new real-world situations, examples, or events.

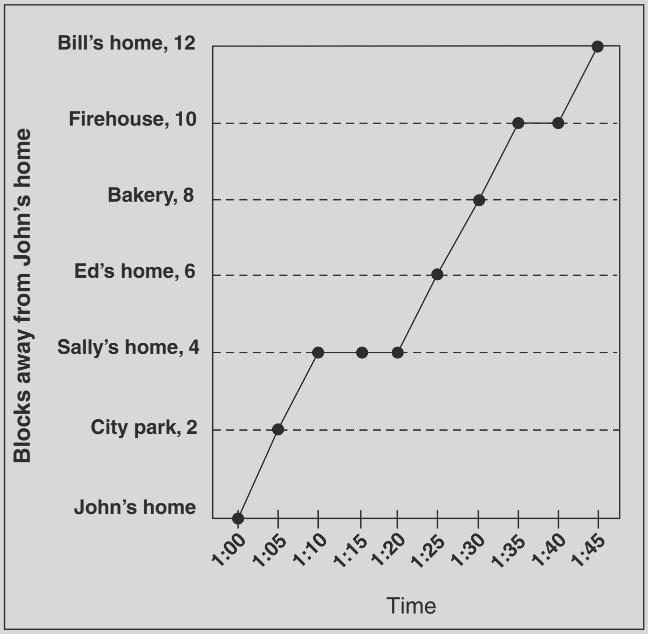
Create the directions last. Be sure the directions clearly describe the basis on which the student is to solve the masterlist item set. For example, in the civics course exercise, the directions tell students they must read the examples in the statements and decide which constitutional principle each represents. In the graph interpretation example, the directions tell the student that the statements are interpretations of the graph and that the student must decide whether the graph supports or contradicts the interpretation. If your masterlist item set refers to interpretive material such as this, your directions should clearly describe the material and how students should use it.

After completing the preceding steps, polish your masterlist item set. Organize the interpretive material, if any, at the beginning of the set. Next, place the masterlist before the stems. *Assign letters to the masterlist response choices* that the students may

FIGURE 9.11 **A masterlist item set that requires students to recognize proper interpretations of a graph.**

Use the graph below to help you answer Questions 1 through 5.

The graph shows that John left his home at 1:00 and arrived at his friend Bill's home at 1:45. The graph shows where John was in the community at different times.



Directions: The numbered statements below tell what different students said about this graph. Read each statement and decide whether the information in the graph is consistent with a student's statement. Mark answer:

A — if the information in the graph **is consistent** with the statement.
B — if the information in the graph **contradicts** the statement.
C — if the information in the graph **neither contradicts nor is consistent** with the statement.

___ 1. John ran or walked very fast between his house and the city park.

___ 2. John stopped at Sally's home on his way to Bill's home.

___ 3. John stopped at Ed's home on his way to Bill's home.

___ 4. John stopped to buy something at the bakery before he got to Bill's home.

___ 5. John traveled faster after he passed by Sally's home than before he reached her home.

use. If you will not use a machine-scorable answer sheet, your letters do not have to be *A, B, C, D,* or *E*. For example, in the civics course exercise, you could use *C* for "consent of the governed," *M* for "majority rules," and so on.

The stems are always numbered and follow the masterlist. If you are not using a separate answer sheet, put a blank *before* the stem number rather than at the end of the stem. You will score the papers much more quickly and accurately if the blank is before the number of the stem. Scramble the stems so that all the stems matching one masterlist option are not together, and make sure there is no discernible pattern to the answers (avoid ABCDABCD, etc.). If you have written too many stems, select the best ones and save the others for revision and use at a later date. Edit the stems to make them grammatically correct, clear, and concise. Limit each stem to 40 or fewer words. However, the stems and the other parts of the item set must provide enough information for the student to apply the rule or principle. For example, in the civics course exercise, Stem 2 would not have had sufficient information if it contained only these words:

1. A civil war was taking place in another country. The president of the United States began planning to help support the civil war.

More details are needed for students to figure out which principle the stem illustrates.

The checklist summarizes the suggestions in this section as a masterlist checklist. Use this checklist to guide you in creating and using masterlist item sets.

✔ **CHECKLIST**

A Checklist for Reviewing the Quality of Masterlist Exercises

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does the masterlist exercise assess an important aspect of the unit's instructional targets?
2. Does the masterlist exercise match your assessment plan in terms of performance, emphasis, and number of points?
3. Does the masterlist exercise require the students to apply their knowledge and skill to new situations, examples, or events?
4. Did you provide enough information so that knowledgeable students are able to apply the knowledge and skill called for by the item?
5. Do your directions to the students clearly and completely explain the basis you intend them to use when applying masterlist response choices to the stems?
6. Within this masterlist exercise, does every stem and every response choice in the masterlist belong to the same category of things?
7. Does every response choice in the masterlist function as a plausible alternative for every stem?
8. Did you avoid "perfect matching"?
9. If possible, are the options in the masterlist ordered in a meaningful way (logically, numerically, alphabetically, etc.)?
10. Are the stems numbered and the masterlist response choices lettered?

Advantages A masterlist item set is a variation of the matching exercise format, and it has many of the same advantages as that format. It is a space-saving and objective way to assess learning targets for which you want students to identify associations between two sets of things. However, it is best used to assess a student's *understanding of concepts*, for example understanding of the constitutional principles in Figure 9.10.

To assess concept understanding, the examples you give students to classify cannot be the same examples you illustrated in class or that appeared in the textbook or assignments. If your examples are not "new to the students," then the masterlist item set becomes simply an alternate way to assess students' recall and recognition of verbal information. As with other matching exercises, you can use pictures, maps, symbols, or diagrams as stems.

The masterlist item also is an efficient way to assess a student's ability to (a) analyze a passage, table, or graph and (b) recognize an appropriate interpretation or conclusion drawn from this interpretative material (see Figure 9.11).

Criticisms Because masterlist item sets are "cousins" to matching exercises, they share the same criticisms. Critics point out that teachers often limit using the format to assess rote associations such as names and dates, memorized lists of causes and effects, lists of symbols and definitions, and so on. Although some learning targets do focus on memorization and recall of information, many do not and should not. As we described, use masterlist item sets to assess students' (a) comprehension of concepts, principles, or schemes for classifying objects, events, and ideas and (b) ability to analyze appropriate interpretations and conclusions.

Tabular (Matrix) Items

A **tabular (or matrix) item** format is a type of matching exercise in which elements from several lists of *responses* (e.g., presidents, political parties, famous firsts, and important events) are matched with elements from a common list of premises. The students' task is to select one or more elements from each response list and match the elements with one of the numbered premises. An example of a tabular item format is shown in Figure 9.12.

You can see that the example is a quadruple matching exercise: (a) match year and president, (b) match year and president's political party, (c) match year and famous first, and (d) match year and important event. One premise list and several response lists that correspond to it can be efficiently organized into a tabular or matrix item format. Notice that each premise is *numbered*. Thus, each premise is scored as a separate item.

First, as always, identify the learning targets you wish to assess. Targets for which students must cross-classify facts or examples, or for which they must identify several characteristics or properties of dates, events, or objects, are most suitable.

FIGURE 9.12 **A tabular or matrix item set.**

Directions: Match the names, political parties, famous firsts, and important events in the columns with the dates in the table below. Write the letter in the proper column in the table. You may use a letter once, more than once, or not at all in any cell in the table.

Presidents	Presidents' political parties	Famous firsts	Important event
A. Coolidge	K. Democrat	N. First airplane flight	U. Atomic bomb on Hiroshima
B. Eisenhower	L. Independent	O. First airplane flight across U.S.	V. Great Depression begins
C. Harding	M. Republican	P. First automobile trip across U.S.	W. NAACP founded
D. Hoover		Q. First telephone talk across U.S.	X. New Deal legislation passed
E. McKinley		R. First transatlantic solo flight	Y. North Pole reached
F. Roosevelt, F.D.		S. First U.S. satellite in space	Z. Panama Canal opened
G. Roosevelt, T.		T. First woman in cabinet	AA . Panama Canal Treaty signed
H. Taft			BB. Social Security Act passed
I. Truman			CC. United Nations founded
J. Wilson			DD. World War I ends
			EE. Korean Conflict begins

	Year	President	President's political party	Famous first	Important event	Score
1.	1901–1904					1. __ ____
2.	1905–1908					2. __ ____
3.	1909–1912					3. __ ____
4.	1913–1916					4. __ ____
5.	1917–1920					5. __ ____
6.	1921–1924					6. __ ____
7.	1925–1928					7. __ ____
8.	1929–1932					8. __ ____
9.	1933–1936					9. __ ____
10.	1937–1940					10. __ ____
11.	1941–1944					11. __ ____
12.	1945–1948					12. __ ____

You should first construct a list of premises. For instance, in the earlier example, the premises were the 4-year time spans defining a U.S. president's term of office. Next, create lists of responses organized into homogeneous groups. You need two or more homogeneous response lists. You should add to each list at least one *plausible* response that does not match any of the premises. This will eliminate the perfect-matching flaw we discussed previously. For instance, in the earlier example, "Eisenhower," "Independent," "First U.S. satellite in space," and "Korean Conflict begins" do not match any of the premises (dates).

Create the table or matrix to correspond to your premise and response lists. Be sure to *number the premises*. Label the columns with the same headings you used for the response lists. For convenience, make a place to record scores at the right of the table, as is shown in the presidential term example.

Directions to students are created next. The directions should clearly tell the students what they are to match, the basis for matching, how they should record their answers, and that a response may be used once, more than once, or not at all.

Create the exercise in a layout modeled after the preceding example. Put the directions at the top and the lists of responses below the directions and above the table. The exercise is easier to understand with this arrangement than when the response lists follow the table. It is also easier for students to read and keep track of the responses if they appear first. Use letters to identify each response. There are fewer student clerical errors if the lettering continues consecutively across the lists as in the presidential term example. Finally, place the table and make places to record scores. A grid is easier for students to use and for you to score.

The checklist for tabular items summarizes the suggestions in this section. Use it as a guideline when creating the tabular item set and evaluating your item set when it is complete.

✔ CHECKLIST

A Checklist for Reviewing the Quality of Tabular (Matrix) Exercises

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does the tabular exercise assess an important aspect of the unit's instructional targets?
2. Does the tabular exercise match your assessment plan in terms of performance, emphasis, and number of points?
3. Do your directions to students clearly explain (a) the basis you intend students to use when matching the responses to the premises, (b) how to mark their answers, and (c) that a response choice may be used once, more than once, or not at all?
4. Do the response choices within each response list all belong to the same category of things?
5. Does every response choice function as a plausible alternative to every premise?
6. Did you avoid "perfect matching"?
7. If possible, are the response choices ordered in a meaningful way (logically, numerically, alphabetically, etc.)?
8. Are the premises numbered and the response choices lettered?
9. On the test page are the directions placed first, the response choices second, and the table third?
10. If possible, is the entire exercise printed on one page rather than split between two pages?

Scoring Scoring is a special concern with the tabular or matrix item format. Two options for scoring are available:

1. *You may score each numbered row as completely correct or incorrect* (score each row as a 1 [completely correct] or a 0 [one or more elements are incorrect]).
2. *You may score each row according to how many elements are correctly placed in its cells* (score each cell in the row as a 1 [correct] or a 0 [incorrect]).

Of these two options, we prefer the second: It gives partial credit and yields more reliable scores.

Special problems may arise when (a) the correct answer requires placing more than one response in a cell but a student enters *fewer or more responses* than should be entered and (b) the correct answer is a blank but a student enters *some response(s)* in that cell. In the presidential term example, for instance, both Roosevelt and Truman were president in the span 1945–1948: Roosevelt died in office while Truman was the vice president. A student may place an *E*, an *H*, or both into the corresponding cell in the "President" column. How should this cell be scored? The correct answer is "E and H," so clearly this should be given full credit or 2 points (1 point for each). Students who mark only *E* or only *H* could be given partial credit (1 point), or they could be given no credit (0 points). Giving partial credit would seem to be the fairest thing to do.

Suppose, however, a student responded with both C and H. Option C is clearly incorrect, but Option H is correct yet incomplete. We recommend giving partial credit (1 point) for the correct portion and not subtracting points for the incorrect Option C. You could make a note to the student that Option C is incorrectly placed, however.

One way a clever student may attempt to "beat the system" is to put the letter of every response option in every cell of the table. Because the correct option(s) would always be included in a cell (along with all the other incorrect options), the student would get 100% if our suggestions for scoring were followed. What should you do if this happens? We suggest that you return the test paper to the student (without penalty) and ask the student to enter in the table cells only those few choices the student believes are correct. You could also alter the directions in future tests to make it clear that you want only a few choices per cell.

Advantages The tabular or matrix item is a useful way to assess whether students can pull together facts and ideas into an organized format such as a table. It is easy to craft when assessing recall of verbal information, such as facts, dates, generalizations, terminology, and characteristics of theories. It is also very efficient when you have one list of premises and many different lists of responses. You may recall that when writing a *basic* matching item, the responses within a list should be homogeneous: that is, all should belong to the same category. When you are writing a basic matching exercise and find the response list becoming heterogeneous, you may wish to reorganize the exercise into a tabular item set.

Criticisms Although it may be possible to create tabular item sets that assess complex or higher-order thinking skills, it is difficult to do so. Most teachers find this format most useful for assessing recall and recognition of verbal information. Because the format is easy to construct, some teachers overuse it (or its cousin, basic matching) and are therefore subject to criticism of focusing on facts rather than problem-solving, critical-thinking, or other higher-order cognitive skills. Also, scoring the set is problematic.

CONCLUSION

This chapter has discussed writing multiple choice items and matching exercises. With the short-answer and true-false items discussed in Chapter 8, these formats comprise the common selected response, objectively scored items types used in paper-and-pencil tests. In Chapter 10, we turn to the essay question, a constructed response format.

EXERCISES

1. For the subject you teach, or one with which you are most familiar, construct one flawless multiple-choice item to assess each of the following abilities. Before writing each item, write a specific learning target that the item will assess. After writing the items, use the checklist for multiple-choice items to evaluate and revise your items. Share your items with the other members of your class.
   a. Ability to discriminate two verbal concepts
   b. Ability to comprehend a principle or rule
   c. Ability to select an appropriate course of action
   d. Ability to interpret new data or new information
2. For the subject you teach, obtain curricular material that has multiple-choice items for teachers and students to use. Select from this material 10 items that would be appropriate for a unit you might teach. Evaluate each item using the multiple-choice checklist. Identify the flaw(s) in each item, and then revise the item to correct the flaw(s). (Be sure to evaluate your revised items so they are flawless.) Discuss your findings with the class.
3. For the subject you teach, obtain curricular material that has matching exercises for teachers and students to use. Select two exercises that would be appropriate for a unit you might teach. Evaluate each exercise using the matching exercise checklist. Identify the flaw(s) in each exercise, and then revise the exercises to correct the flaw(s). (Be sure to evaluate your revised exercises so they are flawless.) Discuss your findings with the class.
4. Evaluate the matching exercise below using the matching exercise checklist. Prepare a list of the flaws found. For each flaw listed, explain why it is a flaw in this exercise. After completing your analysis, revise the exercise so it has no flaws. Share your findings with your class.

Instructions: Match the two columns

A	B
1. chlorophyll	A. Green plants contain this substance
2. igneous	B. Type of rock formed when melted rock hardens

3.	photosynthesis	C.	A substance made up of both hydrogen and oxygen
4.	water	D.	Process by which green plants produce their food

5. Choose two different kinds of relationships from the following list and craft matching exercises for the subject you teach: accomplishments of persons; dates of noted events; definitions of terms; examples of applications of principles or rules; symbols for concepts; authors or artists and their specific works; English equivalents of foreign phrases; functions of specific parts of a mechanism; or names of pictured objects. Develop matching exercises for each of the two you chose. Evaluate and revise your exercises using the matching exercise checklist to eliminate all flaws. Present your matching exercises to your class.

Essay Assessment Tasks

KEY CONCEPTS

1. Essay test items ask students to compose their responses, and are scored with a judgment of the quality of those responses. Restricted-response essay items limit both the content of students' answers and the form of their written responses; extended-response essay items require students to express their own ideas and to organize their answers.
2. Good essay questions ask students to use the higher-order thinking skills specified in their learning targets.
3. Follow item-writing guidelines to create high-quality essay questions that assess subject-matter learning.
4. For summative (graded) assessment, require all students to answer the same essay questions.
5. Create writing prompts to assess writing achievement in different genres: narrative, imaginative, expository, and persuasive.
6. Score essays with rubrics or rating scales.

IMPORTANT TERMS

carryover effect
expository writing
extended-response essay items
halo effect
imaginative writing
independent scoring of essays
narrative writing
optional essay questions
persuasive writing
prewriting activities
prompt
rater drift
restricted-response essay items
scoring reliability
Six + 1 Traits® of Writing
SOAP
writing process
writing traits (writing dimensions)

From Chapter 10 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FORMATS FOR ESSAY ITEMS

Essay formats are usually classified into two groups: restricted-response items and extended-response items. Both types are useful tools, but for different purposes.

Restricted-Response Varieties

Definition **Restricted-response essay items** restrict or limit both the content of students' answers and the form of their written responses. This is done by the way you phrase a restricted-response task. An example follows:

Example

1. Write a brief essay comparing and contrasting the terms *measurement* and *assessment* as they relate to (a) the degree of quantification of the quality of students' responses, (b) the process of obtaining information, and (c) the way in which students' responses are recorded.

Assessing More Than Recall and Comprehension Restricted-response items should require students to apply their skills to solve new problems or to analyze novel situations. One way to do this is to include interpretive material with the assessment. Interpretive material could be, for example, a paragraph or two describing a particular problem or social situation, an extract from a literary work, or a description of a scientific experiment or finding. Essay (and response-choice) items based on this kind of material are called interpretive exercises or context-dependent tasks. Interpretive exercises ask students to read, listen to, analyze, or otherwise interpret the accompanying material and then to complete one or more items based on it. The following examples illustrate restricted-response items and, by way of contrast, an extended-response item that requires students to analyze a particular poem in various ways. The items are intended for a high school literature course.

Example

Interpretive Material

On First Looking Into Chapman's Homer

Much have I travell'd in the realms of gold,
And many goodly states and kingdoms seen;
Round many western islands have I been
Which bards in fealty to Apollo hold.
Oft of one wide expanse have I been told
That deep-brow'd Homer ruled as his demesne;
Yet did I never breathe its pure serene
Till I heard Chapman speak out loud and bold:
Then felt I like some watcher of the skies
When a new planet swims into his ken;
Or like stout Cortez when with eagle eyes
He star'd at the Pacific—and all his men
Look'd at each other with a wild surmise—
Silent, upon a peak in Darien.
—John Keats

Restricted-response questions

1. What is the poet's attitude toward literature as is apparent in lines 1 to 8? What words in these lines make that attitude apparent?
2. Summarize the mood described in lines 9 to 14.
3. What is the relationship between the attitude described in lines 1 to 8 and the mood established in lines 9 to 14?

Extended-response questions

1. Describe the way in which the structure of the poem reinforces the speaker's mood as it is presented in lines 9 to 14. In your essay show how the attitude in the first part of the poem is related to the mood at the end of the poem.

Source: From "Evaluation of Learning in Literature," by A. C. Purves, in *Handbook on Formative and Summative Evaluation of Student Learning* (pp. 736, 755–756), by B. S. Bloom, J. T. Hastings, and G. F. Madaus (Eds.), 1971, New York: McGraw-Hill.

Often restricted-response items are limited to only certain aspects or components of very complex learning. The restricted-response items above, written at the "Analyze" cognitive level of the revised Bloom's taxonomy, ask a few of the many (perhaps 15 or 20) questions that a teacher might write to assess students' ability to analyze the mood of a poem. The extended-response item, by contrast, attempts to elicit from the student a rather complete and integrated analysis of the poem.

Advantages of Restricted Response This format narrows the focus of your assessment to a specific and well-defined performance. The nature of these items makes it more likely that your students will interpret each question the way you intended. You are in a better position to assess the correctness of student answers when a question is focused and all students interpret it in the same way. When you are clear about what makes up correct answers, it improves your scoring reliability, hence the scores' validity.

Other Assessment Options Multiple-choice interpretive exercises assess many abilities more reliably than restricted-response essays. This is illustrated in Figure 10.1. Do not use restricted-response essays only to assess students' recall of factual information. You can assess a student's ability to recall factual information better through completion, true-false, multiple-choice, and matching items.

Extended-Response Varieties

Definition **Extended-response essay items** require students to write essays in which they are free to express and organize their own ideas and the interrelationships among their ideas. There are multiple ways to write a good answer. A student is free to choose the way to respond, and degrees of correctness or merit of a student's response can be judged only by a skilled teacher who is informed on the subject.

Two Purposes The two broad uses for the extended-response essay format are to assess students' (a) general writing ability and (b) subject-matter knowledge. This chapter discusses both of these essay purposes.

Writing Assessment If your intention is to assess only writing ability, your essay must present the students with a prompt. A **prompt** is a brief statement that suggests a topic to write about, provides general guidance to the students, motivates the students to write, and elicits the students' best performance. You evaluate your students' performance by using a scoring rubric that defines various characteristics or qualities of writing.

FIGURE 10.1 Examples of varieties of learning outcomes that can be assessed using objective interpretive exercises and essay items.

Type of test item	Examples of complex learning outcomes that can be measured
Objective interpretive exercises	Ability to— identify cause-effect relationships identify the application of principles identify the relevance of arguments identify tenable hypotheses identify valid conclusions identify unstated assumptions identify the limitations of data identify the adequacy of procedures (and similar outcomes based on the pupil's ability to *select* the answer)
Restricted-response essay questions	Ability to— explain cause-effect relationships describe applications of principles present relevant arguments formulate tenable hypotheses formulate valid conclusions state necessary assumptions describe the limitations of data explain methods and procedures (and similar outcomes based on the pupil's ability to *supply* the answer)
Extended-response essay questions	Ability to— produce, organize, and express ideas integrate learning in different areas create original forms (e.g., designing an experiment) evaluate the worth of ideas

Source: Adapted from *Measurement and Evaluation in Teaching* (7th ed.), p. 224, by R. L. Linn and N. E. Gronlund, 1995. Upper Saddle River, NJ: Prentice Hall. © 1995. Adapted by permission.

Example

Extended-Response Imaginative Prompt

Pretend you are one of the characters in a fairy tale and have just been granted three wishes. What would your first wish be? Write the wish, and then write a story about what happens to you when the wish is granted.

In this example, the prompt stimulates the student to write in an imaginative way. The student is asked to use expressive writing ability to play an imaginative role and write a fantasy narrative. Later in this chapter we will discuss criteria for evaluating students' writing.

Subject-Matter Knowledge Assessment If the primary purpose of your assessment is to evaluate students' knowledge, understanding, and reasoning in a subject, then a different kind of prompt and essay structure is needed. Here is an example in the subject of social studies:

Example

Extended-Response Subject-Matter Essay Prompt

On June 28, 1914, a Serbian nationalist assassinated Archduke Francis Ferdinand, the heir to the Austro-Hungarian throne, in Sarajevo, Bosnia. Describe the political climate that caused this spark to escalate into a war between the Allied Powers and the Central Powers. Your description should explain how these political factors are related to at least two of the more general principles we have studied (e.g., power vacuum), and how those principles operated in what turned out to be the start of World War I.

In this example the prompt set the task for the students by describing the general purpose of the essay the students are expected to develop. Students' responses are evaluated primarily using subject-matter criteria—how well students understand the political factors, how those factors fit with and exemplify political theory, and so forth. The essay is designed to assess students' competence in reasoning and applying knowledge in the subject of social studies.

Advantages Some of your learning targets center around the students' ability to organize ideas, develop a logical argument, discuss evaluations of certain positions or data, communicate thoughts and feelings, or demonstrate original thinking. The restricted-response essay format does not lend itself to assessing these types of learning targets. Students need opportunities for more extended responses to demonstrate such skills and abilities. The extended-response essay is also suited to assessing learning targets that require students to use a combination of skills such as interpreting material, solving a problem, and explaining the problem and its solution coherently.

Disadvantages One disadvantage of an extended-response essay is poor **scoring reliability**. It is difficult to score an extended-response objectively. A common problem is that, without special training, different teachers will award different marks to the same essays. When the grades for the same responses are inconsistent from one teacher to the next, the validity of the assessment results is lowered. Another common problem with teachers' grading of essays is that they evaluate different students' essays using different criteria. For example, you may attend mainly to the quality of ideas in Johnny's paper, to the neatness and grammatical elegance of Sally's, and to the poor spelling in Harry's. As a result of your student-to-student inconsistency, the assessment results are less valid.

A second disadvantage is that scoring essays is often time-consuming, especially if you want to give feedback to students so they can improve their learning. When you have many essays to mark, that leaves little time for giving detailed comments to students on how to improve their work.

A scoring rubric may improve the reliability (consistency), hence the validity, of scoring essays. Using a scoring rubric also reduces scoring time. Chapter 12 gives suggestions for crafting scoring rubrics.

USEFULNESS OF ESSAY ASSESSMENTS

Abilities and Skills Assessed by Essay Items

The preceding paragraphs and Figure 10.1 described some of the abilities and skills that essays let students demonstrate. Notice that multiple-choice items can measure some of these same abilities. Suggestions for developing multiple-choice and essay items to measure higher-level abilities are given in Chapter 11. What is perhaps unique about the essay format is that it offers students the opportunity to display their abilities to write about, to organize, to express, and to explain interrelationships among ideas. You may assess memory, recall, and comprehension more easily with short-answer and response-choice items. Select an assessment format that will assess exactly the learning target you want students to achieve.

Ideas for Phrasing Essay Questions to Assess Different Abilities

You may find it helpful to study different ways of phrasing questions, which will allow you to craft items that encourage students to use higher-level cognitive processes and skills. Figure 10.2 shows some examples of ways to phrase essay questions so they assess different learning targets. Writing essay questions in such a manner will allow you to assess higher-order thinking skills. Notice that many of the questions use interpretive materials that are new or novel for the students. Also notice that most of the questions ask the students to give reasons or explain their choices. Without asking for such explanations or reasons, you will not be assessing the higher-order thinking processes the students use.

Influence on Studying Strategies

You may use assessment to motivate students to study. It seems reasonable that the type of performances you expect from students on tests will influence their methods of study. Some research indicates that when students know that essay questions will be

FIGURE 10.2 **How to phrase essay questions to assess learning targets.**

1. *Concept understanding:* Identifying examples, producing examples
   - Read the newspaper articles attached. Which events illustrate the concept of *political compromise?*
   - Explain in your own words the meaning of *prejudice*. Give an example of prejudice from your own experience.
2. *Concept understanding:* Classifying examples
   - Read the five mathematics word problems attached. Sort these problems into two groups. Explain why the problems in each group are similar and belong together. Explain how the two groups differ.
   - Study the pictures of the 10 paintings that are attached. Organize these paintings into two or more groups according to their style. Explain the reasons behind your grouping.
3. *Analysis*
   - Look at the family photo attached. Describe the mood or feeling in the photo as well as the body language of the people. Use metaphors or similes to make these descriptions.
   - Read the attached newspaper article. Which statements are opinions? Explain why you think so.
4. *Comparison*
   - Compare Artist A's use of color in her paintings with Artist B's use of color in his mask. How are they similar and different? What moods do the colors convey in each piece?
   - Read the attached statements of Senator A and Senator B. In what ways are their points of view similar? Explain the reasons for your conclusions.
5. *Using principles and rules:* Inference, prediction
   - Read the situation above about the Basarwa, a cultural group we did not study. Based on what we did study about cultural groups, what would you predict would happen to the Basarwa in the next 20 years? Explain the principles you used to make your predictions.
   - Suppose the government of South Africa ordered all of the white citrus farmers to leave the country. Where would you expect them to go? Explain the principles you used to make these predictions.
6. *Inferences:* Deductions, predictions, generalizations
   - Compare the information in Table A with the information in Figure A. What conclusions do you draw about how successful rice farming will be in the region to which the data apply? Explain the reasons for your conclusions.
   - Read the attached statements from a scientist, senator, and newspaper editor about the consequences of continuing to use gasoline-powered automobile engines. What generalization can you make about the continued use of these engines in developed countries?
   - Study the data in the table above. What would you expect to happen to our exports of wheat over the next 5 years? Explain the assumptions you made for your predictions to be valid.
7. *Evaluation*
   - Above are the criteria we use to judge how well an author has used "voice" in writing. Attached is a short piece of writing by a student in a nearby school. Use the criteria to evaluate the writer's use of voice. Explain why good voice is or is not used by this writer. Use examples from the piece to illustrate your evaluation.
   - Use your daily log and records of your plant's growth to explain the present state of your plant. Explain why your plant is better or worse than your classmates' plants. What could you have done differently? What effect would that have had on your plant's present state?

asked, they tend to focus on learning broad concepts and on articulating interrelationships, contrasting, comparing, and so on; those preparing for response-choice questions focus on recalling facts, details, and specific ideas (Struyven, Dochy, & Janssens, 2005). But despite reporting that they prepare differently for different types of assessments, students do not necessarily *perform* differently on the different forms.

When a state department of education uses essay questions on its accountability tests, that motivates teachers to require students to write more, and they report that students' writing skills improve (Evaluation Center, 1995). Outside observers report, however, that although students write more, they do not necessarily write better (Viadero, 1995).

Because both essay and response-choice formats can call for knowledge of specific facts, and both can call for application of complex reasoning skills, the questions' format may not be the key issue in how students plan their study strategies. The kinds of study strategies your students use in preparing for your assessments are more likely to reflect the type of thinking skills your assessment tasks require rather than the format (essay or not essay) of the tasks. If two different assessment formats require students to use the same kind of thinking skills, the formats ought to require the same types of study strategies. If your "essays" are really a regurgitation of facts, students' study strategies will focus on remembering and recalling facts. Thus, the advantages of essays and other open-ended response formats will not be realized in your classroom.

If you believe students must learn to write about ideas in a particular subject area, perhaps the best advice is to be sure you explain and teach writing about the subject to students in your class. Assign students a significant number of writing tasks so they can learn to write in this subject area, and do not limit writing tasks to examinations. Various written assignments such as short compositions and longer term papers can help your students achieve these writing-oriented goals. This often means relying less on the questions and homework assignments that are in the back of the students' textbook chapters and more on your own assignments. Keep in mind, however, that assessment results are more valid if you use multiple assessment formats. Your summative assessments should include, therefore, both essay and response-choice items so they cover a proper range of learning targets.

Depth and Breadth of Content Sampling

Answering essay questions takes a long time and limits the breadth of content about which the student can write. If your students can answer one or two response-choice items in 1 minute, then they can answer 30 to 60 response-choice items in a half hour. Sixty items can cover a very broad area of content and at least parts of many instructional objectives. In the same 30 minutes, these same students can probably answer only one or two essay questions. Thus, you can assess in-depth learning of a narrower topic using one essay or broad, less in-depth, general coverage using many objective items. To improve the content coverage of their assessment, many teachers use both essay and objective test items.

To overcome the shortcoming of an essay's limited content sampling, use a series of compositions that students can write over a longer period. You can accumulate these in portfolios. Several out-of-class essays written over a marking period may better assess a particular learning target than a single essay written during a brief examination period. You must remember, too, that asking students to write under the time pressure of an examination may not be the best method to assess their maximum ability.

Efficient Use of Teacher Time

Essays and compositions take a long time to mark properly. This time is well spent if these are the best ways to assess important learning targets, if performing them is a meaningful student activity, and if the students benefit from your feedback on the quality of their responses. Teachers' scoring quality may deteriorate if they must score large numbers of essays. Use essays when they are the most valid form of assessment for a learning target and are worth the time they take to score well.

Influence of Scoring Criteria and Exemplars

Use well-defined criteria to evaluate students' essay responses. Students should be taught the criteria as part of their regular instruction. Because local school districts and state educational authorities have recognized the importance of developing these criteria, teachers may be engaged in professional development activities to help define these criteria and to improve their application to students' responses. Much of this effort is focused on defining criteria or rubrics that are used with performance assessment, of which essay assessment can be considered a part.

Collaboration with other teachers helps teachers craft criteria and select examples of work at different quality levels that clarify the meaning of statements of state standards and of learning targets in the curriculum. Sharing quality criteria and exemplars with students will better integrate your assessment and instruction. Students learn what the characteristics of quality performance are and, through examples, learn what quality performance looks like.

CONSTRUCTING ESSAYS ASSESSING SUBJECT-MATTER LEARNING

The checklist that follows summarizes suggestions for improving essay items. As with previous checklists, an answer of no to any one of the checklist questions is sufficient reason not to use that essay item until you correct the flaw. The suggestions are discussed later in the chapter. First, we will look at a poorly written essay item and apply the checklist to it. This will give you an idea of how the item should be improved.

✔ **Checklist**

A Checklist for Evaluating the Quality of Essays Assessing Subject-Matter Learning

Ask these questions of every item you write. If you answer no to one or more questions, revise the item accordingly.

1. Does the essay assess an important aspect of the unit's instructional targets?
2. Does the essay match your assessment plan in terms of performance, emphasis, and number of points?
3. Does the essay require students to apply their knowledge to a new or novel situation?
4. When viewed in relation to other items on the test, does this item contribute to covering the range of content and thinking skills specified in your assessment plan?
5. Is the prompt focused? Does it define a task with specific directions, rather than leave the assignment so broad that virtually any response can satisfy the question?
6. Is the task defined by the prompt within the level of complexity that is appropriate for the educational maturity of the students?
7. To get a good mark on the item is the student required to demonstrate more than recall of facts, definitions, lists, ideas, generalizations, etc.?
8. Is the prompt worded in a way that leads all students to interpret the assignment in the way you intended?
9. Does the wording of the prompt make clear to students all of the following:
   a. Magnitude or length of the required writing?
   b. Purpose for which they are writing?
   c. Amount of time to be devoted to answering this item?
   d. Basis on which their answers will be evaluated?
10. If the essay prompt asks students to state and support their opinions on controversial matters, does the wording make it clear that the students' assessment will be based on the logic and evidence supporting their arguments, rather than on the actual position taken or opinion stated?

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 31), by A. J. Nitko and T.-C. Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of copyright holders.

Case Study of a Poorly Written Essay Item

Before we discuss the suggestions in the checklist in detail, let's study a poorly worded essay question and use the checklist to evaluate it. This exercise should help you understand how to evaluate your own essay questions and will make the checklist explanations more meaningful to you.

The Poor Item Suppose a teacher wanted to assess the following 10th-grade U.S. history learning target:

Example

Tenth-Grade Learning Target to be Assessed

Analyze reasons for success of the Colonials during the American War of Independence and explain what alternative actions the British or the Colonials could have taken to alter the outcomes.

The teacher wrote the following essay question to assess this learning target. Overall, the teacher's essay item does not assess the learning target very well. Read this item, and then we shall evaluate it using the checklist.

Example

A Poorly Crafted Essay Item

Analyze the defeat of the British by the Colonials by listing the four factors discussed in class that led to the defeat.

Evaluation Using the Checklist Here is a point-by-point analysis using the checklist; the numbers refer to the point in the checklist:

Example

1. Yes, the factors contributing to the success of the Colonials are important to the learning outcome of this unit.
2. No, the learning target calls for students to analyze reasons for success and explain alternative possibilities. The item requires neither analysis nor explanation.
3. No, the item requires only listing (recalling) information presented during the class.
4. Yes, this item, in relation to other items (not shown), contributes to the breadth of coverage the teacher had in mind for the unit.
5. Yes, what the student is to do (i.e., list) is clearly stated.
6. No, the learning target implies that the students should be capable of more than the item requires. (The task set by the item, "listing from memory," is within the capability of the students, but it is below the appropriate level of complexity as specified by the learning target.)
7. No, the item requires only recalling verbal information.
8. Questionable; some students may be confused by the word *analyze* but most will probably make a *list*.
9. a. Yes, the item says students should list four reasons.
   b. Perhaps the purpose is simply to repeat what was taught in class, but the purpose isn't stated.
   c. No, a time limit is not stated.
   d. No, but simply being right or wrong seems to be the implied basis for evaluation.
10. Not applicable; no opinion asked.

The Revised Item After using the checklist, the teacher rethought the item in relation to the learning target and what he had taught. The teacher revised the item to make it more in line with the learning target. Here is the revised item:

Example

An Improved Essay Task

A. List four of the factors that led to the Colonial victory over the British in the War of Independence.	(4 points)
B. For every factor you list, write a short explanation of how that factor helped the Colonists defeat the British.	(4 points)
C. Choose one of these factors that in your opinion the British could have changed or overcome. Explain what actions the British could have taken to change or overcome this factor.	(4 points)
D. What probably would have happened in the war if the British had taken the actions you stated? Why do you think this would have happened?	(8 points)

Grading: Parts A and B will be marked on how correct your answers are. Parts C and D will be marked on how well you support your opinion, but not on what position you take.

Time limit: 40 minutes.

The revised item is more complex and more difficult than the original, but it comes closer to assessing the learning target. Notice that the revised item is expanded to include recalling information, explaining the recalled information, and using higher-level skills. These higher-level skills require students to explain why they hold logically deduced opinions and to describe probable consequences of actions. The teacher's basis for grading is specified, as is a time limit. Because the class period at this school is 50 minutes long, this essay will probably be the only assessment that the teacher could do that day. To cover other aspects of the unit the teacher would need additional assessments, including quizzes, homework, class discussions, and an objective test over the unit's content.

Discussion of the Checklist

1 and 2. *Importance of what is assessed and correspondence to the assessment plan.* We have stressed

that each of your assessment tasks, no matter what their format, must focus on important learning targets and must match your assessment plan. Learning targets that require essays may be difficult for you to state because the target may be complex and abstract. Further, when assessing these complex learning targets, you may need to use more than one type of assessment tasks. For example, you may want a student to demonstrate the ability to analyze critically and evaluate passages expressing different points of view about the equality of men and women. This complex learning target will require assessing the student using several different tasks before you conclude the student has attained this objective.

Focus on the type of response you wish the student to make. You could, for example, write a specimen answer—an outline of the major points you want the students to make. Or you could state the way(s) you expect the student to approach the problem in an essay question. Then you can refine the essay question to clarify what you wish the student to do.

3. *Essential knowledge applied to new situations.* The essay question format has the potential of assessing a student's command of higher cognitive processes and skills. The best way to do this is to require a student to apply thinking skills to new or novel problems and situations. If a student is asked to write only information recalled from the textbook or class discussion, you are assessing only lower cognitive processes. You can better assess recall of information by using short-answer and response-choice formats.

4. *Covering the range of content and thinking skills.* As you read in Chapter 6, your assessment plan should cover your learning targets' full range of content and thinking skills. Your plan plays a key role in guiding your assessment activities. That plan should include using essay questions for complex thinking, and it should balance the available assessment time against the range of coverage you have planned.

5. *Focus questions; clarify limits and purposes.* Phrase each question to focus attention on the issues or points on which you want the students to write. Students will assume the question, as phrased, is exactly what you want them to answer. If they interpret your question in many different ways it will be impossible to evaluate their responses. Consider the extended-response example about analyzing a Keats poem. An unfocused version of this item might read: "Write an essay analyzing the poem." It is unlikely that such an unfocused item would result in an analysis of the poem's "mood," which is what the teacher had in mind. If an item is not focused, you will find it impossible to distinguish those students who can perform the learning target—but misinterpreted your question—from those who simply cannot apply the skills you taught.

Sometimes, if you find it difficult to state the nature of the task itself clearly, specifying the manner and criteria by which you will evaluate students' responses may increase clarity. For example, sometimes a teacher will give students an extract from a newspaper expressing a point of view and want students to evaluate the extracted statement by applying the strategies and criteria taught in class. However, a poorly stated question may simply say, "Do you agree or disagree with this article's position?" There's no telling what kind of responses students would make: Their responses would likely range from a simple yes or no to long-winded polemical entanglements. Focus the item more by specifying which aspects of the extract the students should address and support.

Focusing the question and specifying limits of the intended response do not mean providing information that gives away the answer. If you want the essay to assess the ability to organize a written argument or identify the central issue in a "fuzzy problem," for example, you should not provide students with a particular organization in the question. However, you should tell students that the way they choose to organize the answer is important, and that you will evaluate the paper on how well it is organized.

An important practical suggestion here is to have a colleague or friend review the questions and, if possible, to try the item with a few students. You can then revise the questions if necessary. Following such steps greatly improves the quality of essay questions.

6. *Complexity should be appropriate to educational level.* Because answering an essay requires students to read, think, and write, you must be sure that the item is appropriate to your students' level of educational development. Avoid the use of complicated sentence structures and phrasings for elementary students. Avoid phrases that are indirect or that add unnecessary reading to the question.

Do not, however, oversimplify essays for more advanced students. Essays should challenge students to do their best thinking and use their best writing skills. Because essays require writing, your students must have the level of writing proficiency needed to answer the question. If students do not have sufficient writing skill to express their knowledge on your essay question, you should consider using another means of assessment.

7. *Require more than recall of verbal information.* Although students must learn various facts, ideas, lists, definitions, and generalizations, do not use the essay format to assess this type of learning. Instead, use short-answer, completion, true-false, matching, and multiple-choice formats to assess simple recall of verbal information. These latter formats are better for assessing such recall because they sample more of a student's verbal information store in a fixed time with nonessay items than with essays. Using short-answer and response-choice items increases the content coverage and the validity of the results for assessing recall. Use essays to assess higher-order thinking, including the ability to express one's own ideas, to compare, and explain reasons.

8. *Make the intention of the essay clear.* Make sure your essay question communicates clearly to the students the framework in which they are to respond: the issues their essays are to address; the amount of justification or evidence, the information they are expected to bring to bear in their responses; and the level of detail you expect in their responses.

9. *Clarify response length, purpose, time limits, and evaluation criteria.* You should tell students (a) the approximate length you expect their response to be, (b) the purpose for which they are writing, (c) the goal toward which their essay should aim, and (d) the audience for whom they should target their responses. If you impose time limits, you should clearly announce these to your students. If more than one answer can be correct, your students should know this. If you will deduct points for incorrect spelling, poor language usage, or poor penmanship, tell students before they respond.

10. *Clarify how students' opinions will be evaluated.* Often an essay will require students to state and support their opinions on controversial or nonroutine matters. These essays provide excellent opportunities to assess students' abilities to analyze, synthesize, and evaluate. In such items, you should make clear that students' answers will be evaluated on the logic shown in their answers and how well they use evidence to support their positions. You should reassure them that the opinions or positions they state will not be marked right or wrong per se.

OPTIONAL QUESTIONS

When the purpose for assessment is summative evaluation, you should require all students to answer the same questions. Some teachers believe that offering students **optional essay questions** (a choice of questions) is fairer because it permits students to "put their best foot forward." Research doesn't bear out this belief, however. Some students will choose to answer questions on which they do less well (Wainer & Thissen, 1994). Further, the topics on which questions are based vary in familiarity and difficulty for the students. We have already mentioned how difficult it is to generalize from one essay to the next. In addition, teachers marking essays frequently change their ratings based on their own perceptions of the nature and difficulty of topics. It is extremely difficult, often impossible, to compare tests equitably when students have taken different items (Wang, Wainer, & Thissen, 1995). If all the questions asked on an assessment represent important learning targets, then it seems logical and fair to hold all students accountable for answering all of them.

Perhaps the story would be different if general writing ability were being assessed, rather than subject-matter competence (Coffman, 1971). You could argue that students may demonstrate general writing ability by writing on any one of a number of topics. If you follow this practice, you should score papers on each topic separately, rather than mixing topics together. This will reduce the topic-to-topic differences that tend to raise or lower your rating of an essay quite apart from its merits. As we pointed out earlier in this chapter, however, the topic and the prompt of the essay questions are important determinants of how well a student performs. A student can write well about some topics and poorly about others. You might, for example, write a better essay on the frustrations of a teacher than on the frustrations of a professional golfer, simply because you know more about one area than the other.

Similarly, the topics students choose or are assigned do affect their ability to answer appropriately. Even if you are assessing general writing ability, interpret cautiously students' responses to different topics and prompts. The most important thing for you to do is to use multiple topics and assess students over a period of time, rather than base your evaluation on a single essay.

CONSTRUCTING PROMPTS FOR ASSESSING WRITING ACHIEVEMENT

Assessing students' writing achievement requires special attention to both writing prompts and scoring rubrics. We discuss writing prompts in this section and scoring rubrics for writing assessment in the next section.

Some school district or state assessment programs have adopted very specific writing instruction and assessment frameworks. We cannot discuss all of these in this book. Follow your school's or state's mandated program to be fair to your students. You can adapt the guidelines in this book to your local situation.

General Suggestions for Integrating Writing Assessment and Instruction

Focus on the Characteristics of Good Writing For classroom purposes, teaching and evaluating students' writing should concentrate on characteristics or qualities of good writing, especially those that students can be taught to improve. Sometimes these writing qualities are called **writing traits** or **writing dimensions**. Teaching and assessing writing need to be highly integrated because to improve, students need to know in some detail (a) what dimensions of their writing need improving and (b) how to make these improvements. Information from assessment should allow you to give students specific feedback that guides their writing improvement.

What Are the Characteristics of Good Writing? Educators differ as to what constitutes good student writing. State standards and school district guidelines will differ in the number and type of traits that define good writing. Many schools and states have adopted or adapted some or all of the **Six + 1 Traits® of Writing** developed by the Northwest Regional Educational Laboratory (http://www.nwrel.org/assessment/toolkit98.php). These are:

Example

1. Ideas
2. Organization
3. Voice
4. Word choice
5. Sentence fluency
6. Conventions
7. Presentation

For example, Arizona (http://www.ade.az.gov/standards/6traits/) uses the first six traits (all but *presentation*). Oregon (http://www.ode.state.or.us/teachlearn/testing/scoring/guides/wriscoringguideeng-portrait0609.pdf) uses six traits, plus an additional one, "Citing Sources," for classroom work that requires research.

Organize your assessment *and* teaching around the writing traits you adopt to help students understand what constitutes good writing. Using the traits as a framework for feedback avoids giving feedback that is too general to be helpful (for example, "You need to improve your writing").

Teach Students What Good Writing Is Students need to learn that good writing has certain qualities, and that these traits or qualities are the criteria by which most writing can be evaluated. Students learn that your feedback on how well they have put these traits to work in their writing helps them improve. When the trait framework is made clear to students, you and they will have a shared vision of what good writing is. If this vision is shared across teachers and grades, then students will come to internalize the traits and use them to improve their daily writing.

Integrate Writing Traits With a Clearly Defined Writing Process Part of writing instruction is to teach students that there is an orderly process for developing a piece of writing. All too often, students have the mistaken idea that they should write a final piece at one sitting. This is a far cry from how good writers work. Students need to understand that most writing results from an orderly process that includes drafting, feedback, revisions, and polishing. There is more than one step in this process.

FIGURE 10.3 **How the writing traits may be integrated into a writing process.**

1. *Prewriting activities*—Before writing, a writer clarifies the purpose for writing, begins to organize thoughts, brainstorms, and tries out new ideas. The writer discusses the ideas with others, decides the format and approach to writing, and determines the primary audience. A plan for the piece develops. The teacher may wish to schedule a content conference (Darden, 2000) to help students focus the ideas and content for the piece.
2. *Draft the piece*—The writer works up a preliminary draft of the piece to reflect the prewriting ideas. Ideas and plans change as the draft develops. The purpose for the writing is further clarified (even changed). The draft begins to take shape and ideas and content start to develop. The preliminary organization of the piece is developed so that a beginning, middle, and end begin to emerge. The draft is considered a work in progress, not the final piece.
3. *Obtain feedback for improving the draft*—Based on assessment, the writer gets feedback from the teacher, peers, or others. The assessment is used to make the feedback specific to the traits that have been adopted to define good writing (e.g., ideas, organization, choice of words, use of sentence variety). The teacher may wish to schedule a drafting conference (Darden, 2000) with the student to give some of the feedback.
4. *Revise the piece*—The writer uses specific feedback from the assessment to improve the piece in each of the trait areas. For example, as a result of specific feedback, the writer may incorporate more colorful or more exciting words into a story.
5. *Repeat Steps 3 and 4 if necessary*—The writer may not have implemented the suggestions from the feedback properly. Or, the writer may not understand the feedback and may need more instruction. Writing is not strictly a linear process; it may require many iterations. Student writers must learn that completing the assignment and turning it in is not a final step. The teacher may wish to schedule a process conference (Darden, 2000) with the student to discuss the choices the student made and suggest how to proceed with the revision.
6. *Edit*—The writer edits the revised piece by checking for correct English mechanics: specific points of spelling, grammar, punctuation, etc. English mechanics is one of the traits of good writing. The teacher may evaluate the written piece for how well the student has implemented mechanics. Note that English mechanics are assessed late in the writing process because the pedagogy is to have the student concentrate first on ideas, organization, word choice, etc., as the piece is being developed.
7. *Finalize and make the piece presentable*—The writer puts the piece into final form for presentation to the teacher, with attention to handwriting or word processing, margins, and the like. Attention to appearance is left to the very end, after the piece is revised and polished.

Source: Based on the authors' interpretations of the ideas and suggestions developed at the Northwest Regional Educational Laboratory. Endorsement by the Northwest Regional Educational Laboratory should not be inferred.

The **writing process** presented in Figure 10.3 is adapted from suggestions developed by the Northwest Regional Educational Laboratory. The writing process begins with **prewriting activities** and continues with drafting, assessment and feedback, until a final piece is produced.

Define Standards or Levels of Achievement for Each Writing Trait For assessment-based feedback to be meaningful, you must use standards that clearly identify the student's achievement level on each trait. You can think of achievement as developing along a continuum from very poor achievement at one end to very high-level attainment at the other. The points along this continuum need to be defined so you can pinpoint the students' current level of achievement. Once a student's current level is known, the continuum's definitions of more advanced levels help you guide the student to achieving that next level. Figure 10.4 shows how Oregon's State Department of Education defined the different levels of attainment for the ideas and content trait at the middle school level (http://www.ode.state.or.us/teachlearn/testing/scoring/guides/wriscoringguideeng-portrait0609.pdf). These definitions take the form of a scoring guide.

If your state or school has also adopted a particular framework for writing traits, it probably has also adopted the definitions of different levels of achievement for each of these traits. You will need to use these, rather than craft your own, because students will be expected to write according to them. These descriptions are usually in the form of scoring rubrics or scoring guides.

Rubrics and Trait Definitions Should Apply Across Different Types of Writing Students will be working with different genres or types of writing, as well as writing for different audiences. Because students are novice writers, it is likely to be confusing if each genre and purpose has very different criteria or traits. Pedagogically it is better if the same few traits are applied to many different types of writing. If you and the students evaluate all writing using these same traits (ideas and content, for organization, for word choice, and so on),

FIGURE 10.4 **How different levels of achievement on the Ideas and Content writing trait are defined by the Oregon Department of Education.**

Ideas/Content		
6 **The writing is exceptionally clear, focused, and interesting. It holds the reader's attention throughout. Main ideas stand out and are developed by strong support and rich details suitable to audience and purpose. The writing is characterized by** • clarity, focus, and control. • main idea(s) that stand out. • supporting, relevant, carefully selected details;when appropriate, use of resources provides strong, accurate, credible support. • a thorough, balanced, in-depth explanation / exploration of the topic; the writing makes connections and shares insights. • content and selected details that are well-suited to audience and purpose.	5 **The writing is clear, focused and interesting. It holds the reader's attention. Main ideas stand out and are developed by supporting details suitable to audience and purpose. The writing is characterized by** • clarity, focus, and control. • main idea(s) that stand out. • supporting, relevant, carefully selected details; when appropriate, use of resources provides strong, accurate, credible support. • a thorough, balanced explanation / exploration of the topic; the writing makes connections and shares insights. • content and selected details that are well-suited to audience and purpose.	4 **The writing is clear and focused. The reader can easily understand the main ideas. Support is present, although it may be limited or rather general. The writing is characterized by** • an easily identifiable purpose. • clear main idea(s). • supporting details that are relevant, but may be overly general or limited in places; when appropriate, resources are used to provide accurate support. • a topic that is explored / explained, although developmental details may occasionally be out of balance with the main idea(s); some connections and insights may be present. • content and selected details that are relevant, but perhaps not consistently well-chosen for audience and purpose.
3 **The reader can understand the main ideas, although they may be overly broad or simplistic, and the results may not be effective. Supporting detail is often limited, insubstantial, overly general, or occasionally slightly off-topic. The writing is characterized by** • an easily identifiable purpose and main idea(s). • predictable or overly obvious main ideas or plot; conclusions or main points seem to echo observations heard elsewhere. • support that is attempted, but developmental details that are often limited in scope, uneven, somewhat off-topic, predictable, or overly general. • details that may not be well-grounded in credible resources;they may be based on clichés, stereotypes or questionable sources of information. • difficulties when moving from general observations to specifics.	2 **Main ideas and purpose are somewhat unclear or development is attempted but minimal. The writing is characterized by** • a purpose and main idea(s) that may require extensive inferences by the reader. • minimal development; insufficient details. • irrelevant details that clutter the text. • extensive repetition of detail.	1 **The writing lacks a central idea or purpose. The writing is characterized by** • ideas that are extremely limited or simply unclear. • attempts at development that are minimal or non-existent; the paper is too short to demonstrate the development of an idea.

Source: From *Writing Scoring Guide* (p. 1), by Oregon Department of Education, 1996, Salem, OR: Office of Assessment and Evaluation, author. Reprinted by permission.

students will learn them more quickly and internalize the traits' meanings. As a result, students will more easily apply the traits to all their writing.

Crafting Writing Prompts

Rhetorical Specifications

Students should learn to write for different purposes and audiences and in different genres. To stimulate students to do this, you need to build into your writing prompts rhetorical clues that elicit the kind of writing that you have in mind. The prompts you write should include statements containing the following elements (Albertson, 1998):

1. *Subject*—inform the students whom or what the piece is supposed to be about.
2. *Occasion*—inform the students about the occasion or situation that requires the piece to be written.
3. *Audience*—inform the students whom the intended audience is.
4. *Purpose*—inform the students what the writing purpose is supposed to be: Is it to inform or

narrate? to be imaginative? to be persuasive? (Sometimes the acronym **SOAP** is used for the four preceding elements.)

5. *Writer's role*—inform the students what role they are to play while writing (e.g., a friend, a student, a parent, etc.).
6. *Form*—inform the students if you expect the piece to take a certain form such as a poem, letter, paragraph, essay, and so on.

The following example shows how to improve a writing prompt by adding these rhetorical clues:

Example

Poor: No SOAP—Writing prompt does not provide suggestions for the subject, occasion, audience, or purpose of the piece.

Write a letter telling about an event.

Better: SOAP is built into the prompt

Recall something important that you saw or that happened to you recently. It could be that you saw an accident, a crime, a good deed someone did. Maybe something funny happened to you recently.

Write a letter to a friend to describe what you saw or what happened to you, just the way it happened. Describe the event clearly so your friend who was not there can tell exactly what it was like and how you felt about it.

Writing Prompts for Different Genres

Students should learn to write for different audiences and different purposes. The writing prompts you provide guide them in writing the specific type of piece you have in mind. Typically, classroom writing takes one of four forms: narrative, imaginative, expository, and persuasive.

Narrative writing describes something that really happened, usually a personal experience of a student. Following is an example of a prompt that elicits narrative writing from students:

Example

Narrative Prompt

Think of one HAPPY thing that happened to you in the past. Maybe it was something that happened at home or at school or someplace else.

Write an essay that tells what happened. Be sure to give specific details that explain why this was a happy thing.

Imaginative writing describes something that did not, often could not, happen. Students use imagination and creativity to tell a story. Here is an example:

Example

Imaginative Prompt

Suppose that one day you woke up and found that you were a FISH. What would your life be like? What would happen to you?

Write a story that we can put into our class magazine that tells what happens to you when you are a fish. Be sure to give specific details about what your life as a fish is like.

Expository writing gives an explanation and information. Students are asked to give details, clarify things, and explain things. Here is an example:

Example

Expository Prompt

Animals change a lot when they grow. Think about ONE ANIMAL that you know a lot about.

Write an essay that explains how this animal changes as it grows. Be sure to explain very carefully and clearly so that your classmates reading your explanation can understand.

Students often use expository writing when answering subject-matter essay questions.

Persuasive writing convinces the reader of the writer's point of view. The writer may want the reader to accept his or her idea or to take some actions that the writer supports. Here is an example:

Example

Persuasive Prompt

Suppose students in this school had 30 minutes of free time each week. The school principal wants your suggestions about ONE THING students should do with this free time. What is the one thing you would suggest?

Write an essay to the school principal that would CONVINCE the principal that your idea is the best. Explain why your idea about using the free time is the best and should be followed. Give reasons to support your position.

Additional Suggestions for Writing Prompts

There are some special considerations when preparing classroom assessments that evaluate students' ability to write. Albertson (1998) offers the following suggestions:

Do not prepare prompts that:

- demand specialized knowledge on the part of students.
- ask students to write narratives about experiences that they may not have because of cultural or social background.
- ask for students' opinions about personal values, religious beliefs, or sensitive or controversial matters that parents would object to.
- encourage complaints and criticisms about the school, students' parents, or persons in the community.

Do prepare prompts that:

- refer to specific situations rather than abstract situations.
- will be interesting to students.
- will be interesting to you when you evaluate students' writing.
- are in the realm of the students' experiences.

SCORING ESSAY ASSESSMENTS

Essay questions should be scored with scoring scales that fit the point values planned in the test blueprint (see Chapter 6). Rubrics or rating scales should be used for this purpose. Chapter 12 gives specific details about how to write and apply scoring rubrics. Briefly, rubrics can be categorized in two ways: according to how many scales are used (*analytic* rubrics use several scales; *holistic* rubrics use one) and according to whether the rubrics are *task-specific* or *generic* (or *general*) rubrics.

You may want to go to Chapter 12 now and read the section on rubrics. As an example to have in mind as you read the practical suggestions for scoring essays (below), Figure 10.5 shows two sets of task-specific scoring rubrics for the Keats poem on page 171.

Rubrics have many positive features. Probably the most important is that the descriptions of the qualities of work in general rubrics define what "good work" is and help students conceptualize the kind of performance they are aiming for. The writing trait rubrics shown earlier are an excellent example of this. Thus rubrics are a powerful instructional tool as well as an assessment tool.

Suggestions for Scoring Essays

Principles for scoring essays are summarized in Figure 10.6. We discuss them in the following paragraphs.

Scoring Rubrics Scoring rubrics and model answers were illustrated in the previous example. The point of using these tools is to improve the consistency of your scoring so that you apply the same standards from paper to paper. If your state has adopted general writing rubrics, use them.

Score One Question at a Time If there is more than one essay question, score all students on the first question before moving on. Then grade all answers to the next question. This method improves the uniformity with which you apply scoring standards to each student. It also makes you more familiar with the scoring guide for a given question, and you are less likely to be distracted by responses to other questions. Finally, using this method helps reduce carryover error discussed below. You can reduce carryover errors further by reshuffling the papers after scoring each question.

Score Subject-Matter Correctness Separately From Other Factors When marking subject-matter essays, factors other than an answer's content often affect your evaluation. Among such factors are spelling, handwriting, neatness, and language usage. To avoid blending your judgment of the quality of the ideas or substantive content of a student's answer with these other factors, score the other factors separately—perhaps by using a rating scale (see Chapter 12).

Scoring separately for quality of ideas, correctness of content, and other factors also gives you the freedom to weight each factor appropriately in calculating the grade. For example, you can weight spelling zero or more heavily, depending on the state policy, school policy, or your classroom practice. You still report the results on the zero-weighted factor (e.g., spelling) to the student; you just don't make it part of the grade. But if a factor is to receive a weight of zero, why bother marking and reporting it separately? Two reasons: to allow more complete feedback to students and to allow you to separate your judgment from the substance of the essays, letting you better assess the content learning target.

FIGURE 10.5 **Example of task-specific scoring rubrics.**

The second essay question about our Keats poem read, "Summarize the mood described in lines 9 to 14." First, you must know what a good answer would say. That means you have to understand the poem very well yourself. Chapman did the first good English translations of Homer's *Iliad* and *Odyssey* (which, of course, were written in Greek). At that time (early 1600s), therefore, a whole body of classic literature became available to English-speaking people. This poem is about a reader who reads these works for the first time. He likens literature to a wonderful land ("realms of gold"; lines 1 to 8) and explains that coming across these works of Homer was like discovering a new land. He uses two images: the image of an astronomer discovering a new planet (lines 9–10) and the image of the explorer Cortez discovering the Pacific Ocean (lines 11–14).

Suppose you decided, then, that good student essays would identify these images and conclude that the mood was one of discovery, with its attendant feelings of surprise and delight. You also wanted good essays to be well organized for readers and written according to standard English grammar and usage conventions. These three dimensions (content, organization, and grammar/usage) are your criteria. You might use the following set of rubrics. Note that the content rubric ("description of mood") is task-specific. You could not share this rubric with the students before they wrote their essays because that would analyze the poem for them. Also note that the weights for the content rubric are doubled, making the ideas worth half (6 points) and the writing worth half (6 points).

EXAMPLE OF ANALYTIC SCORING RUBRICS FOR ESSAY QUESTION #2 (PAGE 204)

3 criteria, 12 points possible

Description of Mood (Discovery)

6 Identifies both astronomer and explorer images as discovery images and gives clear explanation

4 Identifies mood but explanation absent or unclear

2 Mood not identified or incorrectly identified

Organization

3 Thesis is clearly stated in topic sentence; how details support thesis is explicitly stated

2 Topic sentence includes thesis; supporting details are present

1 No topic/thesis sentence and/or no supporting details

Grammar/Usage

3 No errors or minor ones that do not impede reading

2 Some errors in grammar or usage, but meaning is clear

1 So many errors that meaning is unclear

Use analytic scoring (above) if feedback on different aspects of performance is required (for example, so a student knows what to work on to improve). Use holistic scoring (below) if one overall judgment is required (for example, on a final exam whose results a student might not see). Notice, however, that the holistic rubrics use the same criteria: content, organization, and grammar/usage. Assign the grade or score whose description most closely matches the student's essay.

EXAMPLE OF HOLISTIC SCORING RUBRICS FOR ESSAY QUESTION #2 (PAGE 204)

A Mood of discovery is clearly identified; support for this is derived from images of astronomer and explorer; writing is clear and well organized.

B Mood of discovery is identified; support is implied but not made explicit in discussion of images of astronomer and explorer; writing is clear and organized.

C Mood of discovery is identified; one of the images is described; organization is minimal; writing needs editing.

D Mood is not clearly identified or is incorrectly identified; writing is neither clear nor well organized.

F Essay is not about mood and/or so many errors in grammar and usage make meaning impossible to interpret.

Notice that your standards of achievement are embodied in these scoring levels. It would be possible to have "harder" or "easier" rubrics, for example, where the D in this scale might be an F in another.

FIGURE 10.6 **Summary of principles for scoring responses to subject-matter essay items.**

1. Prepare some type of scoring guide (e.g., an outline, a rubric, an "ideal" answer, or "specimen" responses from past administrations).
2. Grade all responses to one question before moving on to the next question.
3. Periodically rescore previously scored papers.
4. Score penmanship, general neatness, spelling, use of prescribed format, and English mechanics separately from subject-matter correctness.
5. Score papers without knowing the name of the pupil writing the response.
6. Provide pupils with feedback on the strengths and weaknesses of their responses.
7. When the grading decision is crucial, have two or more readers score the essays independently.

Score Essays Anonymously Scoring is more valid when you do not know the name of the student who wrote the response. Anonymous scoring of essays prevents the halo error described below. Further, if students know that you score papers anonymously, they are likely to perceive the grading process as fair. One suggestion for maintaining anonymity is to have students write their names on the back of the answer sheet or exam booklet. Other, more elaborate methods, such as using student numbers or other codes, are also effective.

Give Students Feedback An important reason for using essays is the opportunity they give you to assess students' expressive abilities and thought processes. You should note strengths and weaknesses in these areas for each student and explain how you arrived at the grade you assigned. Use the suggestions for formative feedback from Chapter 7 to help the essay assessment provide an opportunity for further student learning.

Another suggestion for giving feedback on essays is to hold student conferences—that is, meet with each student individually to review answers and comments. A brief conference of 5 to 10 minutes with each student is more personal and can provide clearer guidance to the student than written comments in the paper's margin. A short, direct conference with each student may also save you hours of writing copious notes and commentary to clarify a point for the student.

Independent Scoring The quirks of individual teachers do affect essay scores. The suggestions in Figure 10.6 help reduce the impact of your idiosyncrasies, but they do not entirely eliminate them. When important decisions rest on the scores from essays, more than one reader is necessary. Realistically, however, even though everyday grading decisions are important, it is unlikely that you will find the time or consistent cooperation of colleagues to carry out **independent scoring of essays**. Nevertheless, such a practice would improve the consistency of your scoring.

Scoring Reliability

The essay format often has very low inter-rater reliability. You can make a deliberate effort to overcome some of the negative factors that lower the reliability of essay scoring. We discuss these factors in the following paragraphs. Attending to these factors will reduce the measurement errors in your evaluations of students' work. You can also improve the inter-rater reliability of essay scores by using scoring rubrics.

Inconsistent Standards Grades assigned to a student's response may vary widely from one reader to the next, both because of the readers' inconsistencies and because of their differences in grading standards. Further, the same reader may mark the same essay differently from one day to the next. The lack of consistent standards in evaluating essays was a major justification for turning to true-false and multiple-choice assessments in education in the early 1900s. A way to overcome this consistency is to have all teachers use the same scoring rubrics.

Rater Drift Even if scoring criteria are well-defined, raters tend either to not pay attention to criteria over time or to interpret them differently as time passes. This tendency to change the way scoring criteria are applied over time occurs slowly and is called **rater drift**. The practical application is that you have to periodically stop and determine whether you are applying the scoring standards the same way to later-scored papers as you did to earlier-scored papers.

Changes in the Topic and Prompt Another factor that causes your assessment results to be inconsistent is the topic (subject) of the essay. A student's scores may vary widely, even when marked by the same reader, because of the topic, prompt, or questions (Dunbar, Koretz, & Hoover, 1991). If you base your evaluation of a student on one essay question, you will not be able to make general statements about this student's performance on different topics. If your statements about a student are limited to only the one essay a student wrote, the validity of your overall evaluation (e.g., grades) is lowered. This is a strong reason for basing a student's marking period grade on multiple assessments collected over the entire marking period according to an assessment plan (see Chapter 6).

Halo Effect The **halo effect** error occurs when your judgments of one characteristic of a person reflect your judgments of other characteristics or your general impression of that person. Thus, you may tend to grade a particular essay more leniently for a student you admire because you know in your heart that the student has command of the objective or topic. The halo effect works the other way, too: You may give a lower grade to a particular essay by a student because you know in your heart that he or she is not a "good student." One way to correct this flaw is to mark essays only after concealing the students' names.

Carryover Effect A **carryover effect** error occurs when your judgment of a student's response to Question 1 affects your judgment of the student's response to Question 2. For example, a student may have a brilliant answer to Question 1 but a mediocre answer to Question 2. The carryover effect occurs when you mark Question 2 right after marking Question 1: You mark Question 2 more favorably because you "carried over" your favorable impression from Question 1. Unless you use the scoring suggestion that follows, the scores you assign to adjacent questions will likely be more similar regardless of the quality of the students' answers than scores on nonadjacent questions. The suggestion is this: Score Question 1 for all students first, then go back and score Question 2 for all, and so on.

CONCLUSION

Essay questions are an important tool for tapping higher-order thinking. Carefully worded and well-scored essay questions can be a window into students' thinking and reasoning with content in any discipline. In the next chapter, we will continue to consider ways to tap higher-order thinking skills. The importance of this can hardly be overstated. After all, when we call someone an "educated person," what we really mean is that he or she can think.

EXERCISES

1. For each subject you teach (or plan to teach), identify different types of material that can accompany context-dependent items.
   a. For each type, state the educational level of the students for which it is intended.
   b. For each thinking-skill category in the examples given in the section titled "Ideas for Phrasing Essay Questions to Assess Different Abilities," write at least one essay item based on the material you identified. Use the examples in the section as models for phrasing your essay prompts.
2. Each of the two essay items that follow has one or more flaws. Using the checklist for improving the quality of essay items, identify the flaw(s), then rewrite each item to eliminate the flaw(s). Check your rewritten essay item to be sure you have not added another flaw.
   a. Item A: State the two examples of prejudices we discussed in class.
   b. Item B: Evaluate the effect of air pollution on the quality of life in the western part of this state.
3. For each essay item you wrote in Exercise 1(b), apply the checklist for improving the quality of essay items. Revise any item for which you answered no to a checklist question. Exchange your items with one or more of the students in this course. Review each other's essay items using the checklist. Discuss with your classmates the reasons for assigning a no to an item. Discuss how to improve each item.
4. Following are four restricted-response essay questions that together constitute a science unit test. After each question is the keyed answer provided by the teacher and Jane Smith's answer. You are to do two things: First, decide the maximum marks (points) of each question. (The entire test has a maximum score of 50 points, so you need to distribute these among the four questions according to what you believe is appropriate.) Second, evaluate Jane Smith's answers against the answer key and award her points according to her answers' degree of correctness.

   Question 1 *What is the shape of a quartz crystal?*
   Answer key: Hexagonal
   Maximum marks: ______
   Jane's answer: "Six-sided hectogon."
   Jane's score: ______

   Question 2 *What is a saturated solution?*
   Answer key: A solution that contains as much dissolved substance as it can for a particular temperature.
   Maximum marks: ______
   Jane's answer: "Large crystals contain a great deal of substance that has been formed. This process of forming crystals is called crystallization. It occurs both in the laboratory and in nature."
   Jane's score: ______

   Question 3 *Write a paragraph describing how you can grow very large crystals.*

Answer key: Any answer that says size of crystal is directly related to the rate of crystallization.
Maximum marks: ______
Jane's answer: "Large crystals contain a great deal of substance that has been formed. This process of forming crystals is called crystallization. It occurs both in the laboratory and in nature."
Jane's score: ______

Question 4 *Name three major categories of rocks*.
Answer key: Igneous, sedimentary, and metamorphic
Maximum marks: ______
Jane's answer: "The three kinds are fire-formed, settled, and those that have changed their form."
Jane's score: ______

5. This exercise should be done during your class.
   a. Compare the maximum marks you assigned to each question in Exercise 4 with those assigned by other persons in this course. (Put the distributions of maximum marks on the board.) For which questions is there more agreement? For which is there less agreement?
   b. Discuss during class the reasons for agreement and disagreement. Make a list of the factors that seem to affect the maximum value that your classmates assign to each question.
   c. Suggest ways of reducing the variability among persons assigning maximum values to questions. Make sure the suggestions are specific to these four questions.
   d. Compare the scores you gave Jane on each question with the scores given by others in this course. On which items is there more agreement? On which is there less agreement?
   e. During class discuss the reasons for an agreement and disagreement in marking. Make a list of the factors that seem to affect the scores assigned to Jane for each question.
   f. Are the questions on which there is more agreement in scoring Jane's responses the same questions on which there is more agreement for maximum marks? Explain.

Higher-Order Thinking, Problem Solving, and Critical Thinking

KEY CONCEPTS

1. To assess higher-order thinking, use tasks that require students to use knowledge or skill in novel situations. Context-dependent items sets are useful for this purpose.
2. A concept is a class or category of similar things. Four strategies for assessing understanding of concrete concepts and four strategies for assessing understanding of defined concepts are presented.
3. A principle is a rule that relates two or more concepts. Four strategies for assessing comprehension and use of rule-governed thinking are presented.
4. Problem solving refers to the kind of thinking required when reaching a goal is not automatic and students must use one or more higher-order thinking processes to do it. Seventeen strategies for assessing problem solving are presented.
5. Critical thinking is reasonable and reflective thinking focused on deciding what to believe or do. Thirteen strategies for assessing critical thinking are presented. Use checklists or rating scales to assess dispositions toward critical thinking.
6. Reading skills involve thinking, too. Three strategies for assessing reading skills are presented.

IMPORTANT TERMS

checklist
closed-response task
cloze reading exercise
concept
concrete concept
context-dependent item sets
critical thinking
defined concept
dispositions toward critical thinking
enhanced multiple-choice items
heuristic
IDEAL problem solver
ill-structured problems
MAZE item type
novel material
open-response task
principle
principle-governed thinking
problem
rating scale
relational concepts
schema (schemata)
well-structured problems

From Chapter 11 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

ASSESSING HIGHER-ORDER THINKING

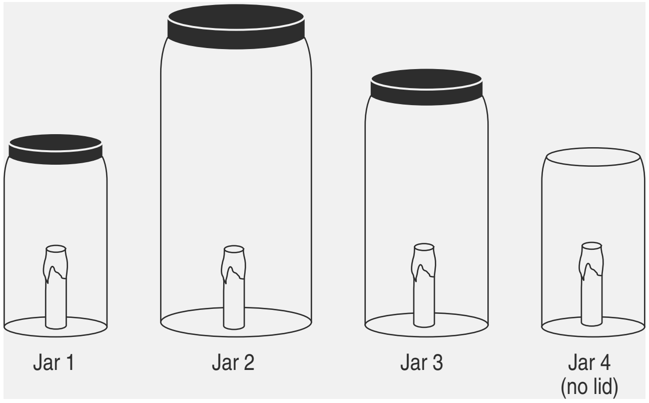
A basic rule for assessment of higher-order thinking skills is to use tasks that require use of knowledge and skill in new or novel situations. If you only assess students' ability to recall what is in the textbook or what you say, you will not know whether they understand or can apply the reasons, explanations, and interpretations. In short, you must use **novel materials** to assess higher-order thinking. One way to do that is to use context-dependent item sets.

Context-Dependent Item Sets

Context-dependent item sets consist of introductory material followed by several items. Students must think about and use the information in the introductory material to answer the questions, solve the problems, or otherwise complete the assessment tasks. Context-dependent item sets are sometimes called interpretive exercises. The introductory material may be extracts from reading materials, pictures, graphs, drawings, paragraphs, poems, formulas, tables of numbers, lists of words or symbols, specimens, maps, films, and sound recordings. Here is one example:

Example

Pat set up four different jars with a burning candle in each jar. He put the lids on jars 1, 2, and 3, as shown in the picture below.



1. The candle in jar 1 burned for 2 minutes after the lid was put on. The candle in jar 2 burned for 8 minutes. About how long did the candle in jar 3 burn after the lid was put on?
   A 1 minute
   *B 4 minutes
   C 8 minutes
   D 10 minutes

2. Pat did not put a lid on jar 4. The candle in jar 4 burned for a very long time. Tell why this candle kept burning so much longer than the other candles.
   ____
   ____
   ____
   ____
   ____

Source: National Assessment of Educational Progress released item Science Grade: 4, Block: 2005-4S12 No.: 3–4.

In this example, the interpretive material is a diagram of a science demonstration. This extract "simulates" a classroom laboratory exercise and thus presents a concrete, realistic example. A student must analyze or process the material in this example to answer the questions. The example shows a multiple-choice item and a short constructed-response item. Context-dependent item sets may be used, however, with any type of item format.

Ability to Use Reference Materials

Assessing the ability to use reference materials, maps, graphs, and tables also lends itself to using context-dependent item sets. This is true whether you are assessing the ability to use both general reference materials or special subject-matter-specific materials. Reference-using skills you may teach and assess include: alphabetizing, using tables of contents and indexes, using encyclopedias, using dictionaries, using general reference materials (calendars, maps and globes, textbooks, periodical indexes, atlases, and so on), using library services, and using the Internet and computer-based CDs. Skills in using these media should also be taught and assessed.

In this assessment area, interpretive materials might include a section of an index, a section of a table of contents, a part of an atlas, a picture of a computer screen, and the like. You may have to rewrite or modify these materials before they are suitable for use in assessment, because (a) they contain material irrelevant or extraneous to assessing the objective at hand; (b) they are too long; or (c) the extract is out of context and is therefore not clear to students. You may need to obtain written permission to reproduce copyrighted materials. You may, of course, use entire volumes or take students to the library for the assessment. To do so, you will need sufficient materials (or

FIGURE 11.1 Examples of items written to assess graph and table reading skills.

Use the table below to answer Questions 1 and 2.

Average Temperature and Rainfall at Windy Hill Town

	2000		*2001*		*2002*		*2003*	
	Temp	**Rain**	**Temp**	**Rain**	**Temp**	**Rain**	**Temp**	**Rain**
September	64°	0.1 in	63°	0.2 in	66°	0.0 in	64°	0.3 in
October	72°	0.4 in	71°	0.5 in	74°	0.4 in	71°	0.6 in
November	77°	0.9 in	75°	1.0 in	78°	0.8 in	76°	0.7 in
December	81°	2.0 in	80°	2.7 in	85°	1.5 in	80°	2.1 in

Example of assessing the ability to locate and compare information from a table

1. When did the highest average rainfall occur?
   A November of 2000
   B November of 2001
   *C December of 2001
   D December of 2003

Example of assessing the ability to draw inferences based on trends and other information in a table

2. Which of the following events was most likely to have occurred between September and December of 2002?
   A The roads were covered with ice and snow.
   *B The town's water reserves were very low.
   C The river flowing through the town overflowed its banks.

computers) for all students, as well as sufficient uninterrupted time to administer this type of performance assessment.

Graphs and Tables

Much information is condensed in tables and graphs. Graph and table reading abilities are important to further learning in many areas, both in and out of school. Examples of some of the graph- and table-reading abilities that you can teach and assess include comprehending the topic on which a table or graph gives information, recognizing what is shown by each part of a graph or table, reading amounts, comparing two or more values, and interpreting relationships, trends, and other main points from the graph or table.

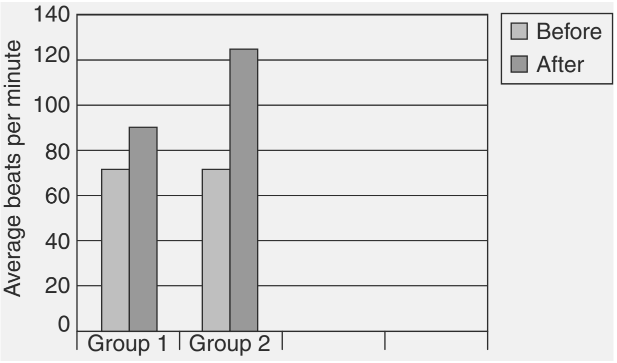
Item 1 of Figure 11.1 requires a student to read the table and locate the information in a cell and to compare several values read from the table to determine which is largest. Item 2 requires a student to make an inference concerning the likelihood of an event based on understanding the trends and facts presented.

Here is an example of how you could use a graph and multiple-choice items to assess capabilities to draw inference based on the displayed rates or trends, underlying relationships, and facts:

Example

Use the Information below to Answer Questions 1 and 2

Before the exercise period began, the teacher divided the class into two groups. Group 1 was to walk around the track two times. Group 2 was to run around the track one time. All students took their pulses both before and after going around the track. The average pulse for each group is shown in the graph below.



Example of an item to assess the ability to draw an inference from a graph

1. According to the graph, which type of exercise made students' hearts beat faster?
   *A Running
   B Walking
   C Neither—they had the same result with either walking or running

Example of an item assessing the ability to interpret trends underlying a graph

2. What would be the heartbeats about 1 hour after the exercise period when all the students are reading in the library?
   *A About 70 for both groups
   B About 70 for the group that walked twice around and about 130 for the group that ran once around
   C About 90 for the group that walked twice around and about 130 for the group that ran once around
   D Lower than 60 for both groups

Maps

Context-dependent items sets are useful for assessing map-reading ability, as well. Specific map-reading abilities include orienting maps and determining direction, locating and/or describing places on maps and globes, determining distances, tracing routes of travel, and interpreting time zones, landscapes, features, and the like. Below is an example of an item that assesses map-reading ability.

Example

(*Population density map of the United States goes here*)

3. A megalopolis is defined as a "supercity" made up of large cities with highly populated areas between them. Look at a population density map of the United States. Which pair of cities is part of a megalopolis?
   A Denver and Salt Lake City
   B Oklahoma City and Dallas
   *C Boston and New York City
   D Kansas City and St. Louis

Source: National Assessment of Educational Progress, released items, Grade 4 Geography, Block: 2001-4G7, No.: 3.

Advantages and Disadvantages of Context-Dependent Items

Advantages A context-dependent item set has these advantages: (a) It provides an opportunity to assess students on materials that are relatively close to the real-world contexts; (b) it provides, through the introductory material, the same context for all students; (c) its introductory material lessens the burden of memorizing and may moderate the effects of prior experience with the specific content; and (d) frequently, it is the only means to test certain intellectual abilities.

Disadvantages Some disadvantages of a context-dependent item set are (a) the set may be difficult to construct, (b) you must carefully create the introductory material to assess higher-order thinking skills, (c) a student's performance on one context-dependent item set may not generalize well to performance on another similar set, (d) the set often requires students to use additional abilities (such as reading comprehension and writing skills) that may go beyond the major focus of the assessment tasks, and (e) you may need special facilities (such as a photocopy machine and/or drawing skill and equipment) to produce them that are not readily available.

Layout

The way context-dependent material is arranged on the pages of a test booklet is important because a poor arrangement may cause students to misread or misinterpret the item set. A side heading and directions should point students to the introductory material and to the particular tasks based on it. The introductory material is placed in the center of the page with items below it.

Keep the introductory material and all items that refer to it on the same page, if possible. Otherwise, students will be distracted as they flip pagesback and forth while completing the assessment. Students with poor short-term attention and memory may lose their place or make careless errors.

CONCEPT LEARNING

What Are Concepts?

A **concept** is a class or category of similar things (objects, people, events, or relations). Many of the things you teach are concepts. Students' understanding of concepts forms the basis for their higher-order learning. When we speak of the concept *red,* for example, we refer to a category of objects with a similar color. A student is said to have learned the concept *red* if the student (a) can identify examples or instances of red things (red tricycle, red book, red lipstick, etc.) *and* (b) does not refer to things that are not red (green tricycle, purple book, pink lipstick, etc.) as red. Concepts are ideas or abstractions: Only specific examples of a concept exist in the world. The individual members of the concept category are called *instances, examples*, or *exemplars*.

A distinction can be made between concrete concepts and defined concepts (Gagné, 1970). A **concrete concept** refers to a class, the members of which have in common one or more physical, tangible qualities that can be heard, seen, tasted, felt, or smelled. Examples of concrete concepts include *large, triangle, green, house,* and *dog*. A **defined concept** refers to a class for which members can be defined in the same way by attributes that are not tangible. Defined concepts frequently involve relationships among other concepts and are sometimes called abstract or **relational concepts** (Gagné, 1970). Defined concepts are usually learned by definitions. Gagné gives the example of *diagonal,* which is defined as a line connecting the opposite corners of a quadrilateral figure. The relationship is "connecting." The related concepts are "opposite corners," "quadrilateral figure," and "line." Other examples of defined concepts include *beside, friendliness, uncle,* and *mother*. Some concepts are learned initially as concrete concepts and later as defined concepts.

Understanding a concept goes beyond simply identifying examples of it. Concepts are related to each other and linked together in complex ways through schemata or networks. A **schema** is the way knowledge is represented in our minds through networks of connected concepts, information, rules, problem-solving strategies, and conditions for actions. For example, Woolfolk (2005) points out that we know counterfeit money is not real, even though it fits the *money* concept prototype and examples. We know it is counterfeit money because we link our concept of money to other concepts, such as the concepts of authority to print, crime, forgery, and so on. You need to help students connect concepts to their existing networks and schemata of knowledge before they can fully understand these concepts.

Strategies for Assessing Concrete Concept Learning

Figure 11.2 presents four commonly used strategies that can be used to assess whether a student has learned a concrete concept. These include students (1) naming the concept from examples, (2) discriminating examples from nonexamples, (3) producing their own examples, and (4) using the concept in performance assessment.

Give the Name Strategy 1 is usually an unsatisfactory way to assess concept learning. Students may learn the concept and perhaps can use it without learning the proper name of the concept. The give-the-name assessment strategy does not require students to discriminate the exemplars from nonexemplars, so you do not know whether the students have overgeneralized the concept. For example, students may think all shapes with round edges are circles. As a result, students may confuse circles with ellipses (ovals) and spheres (balls). Finally, this assessment strategy does not require students to use or apply their understanding of the concept. Thus, even though students can state the concept name, you do not know whether they have the deeper understanding of the concept necessary to connect it to other concepts and integrate it into their schemata.

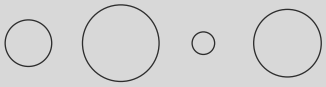
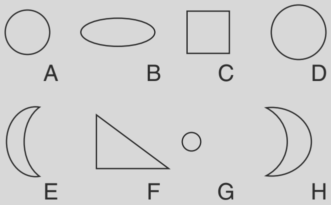
Discriminate Exemplars From Nonexemplars Strategy 2 requires students to discriminate circles from other shapes. This assessment strategy is preferred over that of Strategy 1 because it does not require students to produce the concept name to complete the task and allows you to control the assessment situation. You need to control (a) the degree to which the exemplars and nonexemplars are familiar to students, (b) how typical the exemplars are of the concept, (c) the number and type of discriminations between exemplars and nonexemplars, and (d) the total number of exemplars you present. To use this strategy you must present at least two exemplars for students to identify; otherwise you do not know whether the students have undergeneralized the concept. Undergeneralizing means that the students think only one example is the same as the whole concept (e.g., thinking that the circle you showed in class on the board is the only circle). As with Strategy 1, this strategy does not require students to use or apply their understanding of a concept. Thus, it does not permit you to assess students' deeper understanding of it.

Produce New Exemplars Strategy 3 requires students to think up examples and you to judge the correctness of the examples. You also know whether the students' examples were explicitly taught, in which case you are assessing only remembering. This strategy may be useful for assessing simple concepts (such as circle), but it is not preferred for more complex concept assessment.

Performance Assessment for Deeper Understanding None of the first three strategies assess students' deeper understanding of a concept. Students show their deeper understanding when they are able to (a) use the concept to solve problems; (b) relate the concept to other concepts, principles (rules), and generalizations they have learned; and (c) use the concept to learn new material. To assess students' deeper understanding, you must create assessments that are more complex and require more application of the concept than was illustrated by the previous items.

The example for Strategy 4 in Figure 11.3 shows a *performance task,* meaning that students have to do

FIGURE 11.2 Strategies for assessing concrete concept learning.

Strategy for assessing concrete concept learning	Example
1. *Have students name the concept after seeing exemplars.*	1. What are the shapes in this group called?  [Ans.: Circles]
2. *Have students discriminate concept exemplars from nonexemplars.*	2. Which of these shapes are circles?  [Ans.: A, D, G]
3. *Have students produce their own exemplars when given the concept name.*	3. Draw three circles. Be sure each is different from the others.
4. *Use performance assessment to assess concept understanding at a deeper level*	4. *The teacher explains the task orally* We have been studying two shapes: circles and squares. Yesterday we walked through the school neighborhood to look at the shapes of the buildings, people, and cars. Today, you will draw a picture of the school neighborhood, your neighborhood, or a city. Your drawing must include buildings, people, and cars. However, you cannot use any circles or squares in your drawing. As you work, ask a friend to keep checking your picture to see if you have used either of these shapes. If you have, change your drawing so it has no circles or squares. When you are finished I will look to see if you have used either of the shapes. While you are working, I will ask you to explain to me, using the words for the shapes we have studied, what you have learned in this assignment about the importance of shapes in our world. I will ask you to explain to me what makes drawing a picture like this without circles and squares so difficult. Before you begin I will show you some examples of pictures, explain what I will be looking for and the kinds of answers I will be expecting of you.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (pp. 61–62), by R. J. Marzano, D. Pickering, and J. McTighe, 1993, Alexandria, VA: Association for Supervision and Curriculum Development. Reproduced with permission of McREL, 4601 DTC Boulevard 500, Denver, CO 80237. Telephone (303) 337-0990. © 1993 by McREL Institute. All rights reserved.

something with their knowledge. This assessment can be used to evaluate the following learning:

I. Content targets (in combination)
   A. Identifies circles and squares
   B. Discriminates circles and squares from other shapes
   C. Understands the importance of shapes in the world

II. Complex thinking target (problem solving)
   A. Identifies things that keep you from solving a problem

The assessment in this example takes a relatively long time to administer, probably two or three 40-minute mathematics periods. The main performance involves the students creating a drawing of the neighborhood that includes buildings, cars, and people but that does not use circles and squares. This drawing presents a problem to be solved: How can you depict buildings, cars, and people, yet not use circles and squares? This is a difficult problem for a first grader because circles and squares are basic shapes comprising much of the students' experience. Students must distinguish among circles, squares, and other shapes to solve the problem. Notice, too, that the assessment requires you to do more than collect the drawings. You must interview or have a conference with each student using the drawing to prompt or draw out from the student information about how well the learning targets have been attained.

Figure 11.3 gives an example of rubrics you could use to evaluate students. All learning targets are represented in the scoring rubrics. However, you do not treat each rubric as a separate test item. You assess this task overall, rather than piece by piece, using a combination of activities including reviewing a student's drawing, conferencing with a student, and prompting a discussion of the mathematical content to obtain lots of information about the student's achievement of the learning targets. Only then do you evaluate how well the student has learned each target.

An advantage of using Strategy 4 for assessing concepts through complex performance tasks, such as the one in the previous example, is that students use the concepts in realistic situations. These situations activate students' cognitive frameworks and schemata. They require students to link the concepts to many other concepts as they complete the task. If you focus on a student's way(s) of using the target concepts while he or she engages in problem solving, you assess whether the student understands the concepts beyond simply naming and identifying them.

Strategies for Assessing Defined Concept Learning

Figure 11.4 presents four strategies that can help you assess students' learning of abstract or defined concepts. Of the four strategies, Strategy 1, requiring students to produce a definition, and Strategy 2, requiring students to produce new exemplars of the concept, are the weaker strategies and may not be suitable for younger students. Strategy 3, requiring students to discriminate exemplars from nonexemplars, and Strategy 4, requiring students to identify components and demonstrate relationships, are the stronger strategies. Their main advantage is that they require students to recognize new exemplars, ensuring that they do not respond with rote memorization of definitions. The performance (drawing and labeling) aspect of Item 5 (Strategy 4) has the additional advantage of not depending solely on highly developed verbal skills.

You cannot assess a student's comprehension of some concepts using the types of items shown in the preceding examples. Two of these are (1) relational concepts (e.g., uncle, aunt) and (2) concepts whose exemplars can be described verbally only by repeating the concept name for each exemplar (Anderson, 1972). An *aunt* is a sister of a mother or father. If you tried to write an "instance" of *aunt,* you would need to mention this relationship in the options.

The concept *wings* is an example of the second type of concept (Anderson, 1972). Each exemplar you write would have to include the word *wings* (airplane wings, bird wings, angel wings, etc.), and so a test item would be answerable on the basis of matching a word in the stem with a word in the options. However, you can assess a concept such as *tools* by the types of items shown in the examples because instances of tools (screwdriver, wrench, saw, etc.) can be written without repeating the term *tool.*

ASSESSING WHETHER STUDENTS' THINKING USES RULES

Another important area of learning is rule-governed or principle-governed thinking. A **principle** is a rule that relates two or more concepts. Students learn abstract principles in later elementary and high school. Following are some examples:

FIGURE 11.3 Scoring rubrics for the circle and squares drawing problem.

Identifies and discriminates circles and squares

3 Identifies circles and squares with little or no prompting.

2 Sometimes confuses circles and squares with ellipses or other curves; sometimes confuses squares with rectangles or other shapes.

1 Demonstrates severe misunderstanding of circles and/or squares.

Understands importance of circles and squares as the basic shapes comprising objects in the world

3 Demonstrates a thorough understanding of how circles and squares are the basic shapes that make up most objects in the world.

2 Displays an incomplete understanding of how circles and squares are used in the world and has some notable misconceptions about their use.

1 Does not understand how circles and squares are used in the world.

Understands that not being able to use circles and squares is an obstacle to depicting real-world objects accurately

3 Accurately identifies the most important obstacles or constraints imposed by not being able to use circles and squares in drawings of objects.

2 Identifies some constraints or obstacles about not using circles and squares that are accurate but also includes some that are inaccurate or irrelevant to the drawing problem.

1 Does not identify the most significant constraints or obstacles imposed by not being able to use circles and squares to solve the drawing problem.

FIGURE 11.4 Strategies for assessing defined concept learning.

Strategy for assessing defined concept learning	Example
1. *Have students produce a definition.*	1. Define a *prejudiced act* in your own words.
	2. Tell what is meant by *lonesome*.
2. *Have students produce examples.*	3. Describe two examples of *acts of prejudice* that were not discussed in class or in the text but which you witnessed or experienced during the past few weeks.
3. *Have students discriminate exemplars from nonexemplars.*	4. Which statement *most nearly* describes the concept of *lonesome*? A Ten-year-old Meghan decides to play alone today with her dollhouse, even though her friends asked her to play with them. B Each morning Professor Cory closes her office door to be by herself to write up her research reports. *C Each lunch period 15-year-old Marya stands by herself, not speaking to anyone in the crowded school cafeteria. D Clarisse, a cloistered nun, speaks to no one and spends many hours alone while praying.
4. *Have students identify components and demonstrate relationships.*	*(Picture of earth with person on it omitted to save space)* 5. In the picture above, draw lines and an angle (or angles) to show the location of the zenith in relation to the person. Label the angle(s) and the zenith.

Source: Item 5 is based on ideas in Gagné & Briggs, 1979, p. 227.

Example

Abstract principles learned in high school

- When performance is followed by a reinforcing event the probability of that performance reoccurring increases.
- Experimental studies allow conclusions regarding functional relations while correlational studies allow only statements of co-occurrence.
- People tend to migrate to, and find success in, physical environments closely resembling those from which they came.

- The status of a group in a society is positively related to the priorities of that society.
- The rate of increase in law enforcement officials is negatively related to the stability of the society.

We say students use **principle-governed thinking** when they can apply a principle or rule appropriately in a variety of "new" situations. You must assess students' understanding of a principle by asking them to apply it to a new situation rather than by simply mimicking your classroom.

Strategies for Assessing Comprehension of Rules and Principles

Most principles operate under certain conditions and not under others. Further, when the conditions exist and when a principle does operate, it leads to certain consequences and not to others. This suggests four basic strategies (Figure 11.5) for creating tasks to assess students' comprehension of principles.

These tasks are difficult, especially for younger, inexperienced learners who are not well read. Further, students' performance on such tasks may be difficult to interpret. Here are some of the questions you have to answer about the students' responses to evaluate them properly:

- Are the students' examples new, or were they presented in the class or in the assigned materials?
- Why can't students give good examples?
- If students cannot write an explanation, do they understand the principle?
- Is there weak knowledge of the specific content to which you have asked the principle to be applied?

This type of item requires students to recall the principle without prompting and articulate it. Students unable to do these two things will not answer correctly. Further, there may be more than one correct explanation for the phenomena stated in your example. This occurs often when the "truth" of the principle or its applicability to all situations is open to question.

Because the items implementing Strategies 1 through 4 are highly verbal, they are likely to require a good level of reading comprehension. Students with poor reading skills who actually understand the principle may miss the item. You may try reading the item situations to poor readers

FIGURE 11.5 Strategies for assessing rule-governed thinking.

Strategy for assessing rule-governed thinking	Example
1. *Have students produce or identify consequences.*	1. Suppose the federal government increased the prime lending rate by one-half percent tomorrow. What would you expect to happen?
2. *Have students produce the consequences and explain why.*	2. Suppose hard economic times forced many cattle ranchers from the high plains of the United States to leave the country. Name two or more geographical locations in the world you would expect them to move. Explain your choices.
3. *Have students produce an explanation only.*	3. In the 2004 presidential election between George W. Bush and John Kerry, voter turnout was high. Why do you think this was the case? In your explanation, apply what you know about voter turnout to the 2004 current events context.
4. *Have students draw a conclusion based on application of a principle.*	4. A farmer planted a cornfield, then divided it into two halves. Both halves were planted with the same amount of the same kind of corn and received the same amount of water and sunlight. The farmer tossed a coin to decide which half would receive a new fertilizer and which half would be fertilized with the same product as usual. Corn yield on the half-field with the new fertilizer was 25% more than for the other half-field. What can the farmer conclude?

to see if they will respond better. You also may be able to simplify the reading level.

PROBLEM SOLVING

The Nature of Problem Solving

Students incur a **problem** when they want to reach a specific outcome or goal but do not automatically recognize the proper path or solution to use to reach it. The problem to solve is how to reach the desired goal. When students cannot automatically recognize the proper way to reach the desired goal, they must use one or more higher-order thinking processes. These thinking processes are called *problem solving*. If the procedure for attaining a goal is so well known to students that they can complete the task without having to reason, they do not have to use problem-solving skills. Older students have a name for these kinds of tasks: They call them "no-brainers." They recognize that there is no problem to solve if you do not have to think about the proper solution.

This intuitive concept, or no-brainer, should be a useful clue when you craft tasks to assess problem-solving ability. If the tasks require students simply to repeat a procedure you taught them in a situation that is more or less identical to the one you used in class, you have created a no-brainer task, and not a problem-solving task. To apply problem-solving skills, students need a task that is somehow different or new to them. The task need not be new to the world, just new to the students.

Well-Structured and Ill-Structured Problems

Most of the problem tasks in teachers' editions of textbooks and in the end-of-chapter exercises in student texts are a few notches above no-brainers. They present tasks that are clearly laid out: All the information students need is given, the situations are very much the same as you have taught in class, and there is usually one correct answer that students can reach by applying a procedure you taught. These are known as **well-structured problems** (Frederiksen, 1984). Well-structured problems serve a useful purpose in giving students opportunities to rehearse the procedures or algorithms you taught.

However, well-structured problems are unlike the real-life or authentic problems students will eventually have to face. Most authentic problems are **ill-structured problems** (Simon, 1973). For ill-structured problems, students must (a) organize the information to understand it; (b) clarify the problem itself; (c) obtain all the information needed, which may not be immediately available; and (d) recognize that there may be several equally correct answers. A problem with a single correct answer is called a **closed-response task**; a problem with multiple correct answers is called an **open-response task** (see Collis, 1991).

General versus Subject-Specific Problem Solving

Controversies still exist among cognitive scientists, psychologists, and educators concerning whether we should teach students problem-solving strategies that are general or specific to each curriculum area—strategies specific, for example, to mathematics, history, or art. Strategies for solving curriculum-specific problems are less applicable across different subjects but more powerful within the specific curriculum; the general approach applies somewhat to every curriculum area but has limited power within any specific curriculum.

It appears that people actually use both general and specific strategies (Alexander, 1992; Perkins & Salomon, 1989; Shuell, 1990). Persons working in an area who have a great deal of knowledge and expertise apply well-known problem-solving strategies to solve problems specific to their area. However, if they work outside their area of expertise, the specific strategies no longer apply: They resort, then, to more general problem-solving strategies. However, as they develop expertise in an initially unfamiliar area, the general strategies are dropped in favor of more area-specific strategies.

Heuristics for Solving Problems

Knowledge-based problem-solving methods within a particular domain provide much better solution strategies than the general methods suggested in this section (Anderson, 1987; Royer, Cisero, & Carlo, 1993). Nevertheless, when students do not have a knowledge-based strategy, a heuristic should be tried. A **heuristic** is a general problem-solving strategy that may help solve a given problem. The following is a list of 10 problem-solving heuristics (Cyert, 1980; Frederiksen, 1984):

1. Try to see the whole picture; do not focus only on details.
2. Withhold your judgment; do not rush to a solution too quickly.
3. Create a model for a problem using pictures, sketches, diagrams, graphs, equations, or symbols.

4. If one way of modeling or representing the problem does not work, try another way.
5. State the problem as a question; change the question if the original does not suggest a solution.
6. Be flexible: Look for unconventional or new ways to use the available tools; see the conventional in new ways; try responding to the situation from a different angle or point of view; think divergently.
7. Try working backward by starting with the goal and going backward to find the solution strategy.
8. Keep track of your partial solutions so you can come back to them and resume where you left off.
9. Use analogical thinking: Ask, "What is this problem like? Where have I seen something similar to this?"
10. Talk about and through a problem; keep talking about it until a solution suggests itself.

The IDEAL Problem Solver General problem-solving skills may be organized into a five-stage process that Bransford and Stein (1984) call the **IDEAL problem solver**:

I Identify the problem
D Define and represent the problem
E Explore possible strategies
A Act on the strategies
L Look back and evaluate the effects of your activities

Strategies forAssessing Problem-Solving Skills

Because the more powerful problem-solving strategies are specific to a domain or subject matter, it is difficult within space permitted in this book to present detailed examples. Further, the variety of problems within a curriculum area is quite large, so even a sample of problems may not do justice to the subject. For example, in junior high school mathematics you could craft many problems in content areas such as number and operations, patterns, prealgebra, geometry and measurement, and data analysis (Lane, Parke, & Moskal, 1992).

If you evaluate only whether an answer is correct or incorrect, you are likely to miss the opportunity to evaluate students' thinking skills in general and problem-solving skills in particular. Assessing students' problem-solving skills requires set tasks that allow you to systematically evaluate students' thinking about problem solving. You need to craft different types of tasks to assess the different aspects of problem solving.

Figure 11.6 shows assessment strategies grouped according to the IDEAL problem solver categories. The strategies suggest the general layout or structure of the tasks. You should apply them specifically to your own teaching area.

CRITICAL THINKING

Curriculum frameworks frequently state that developing students' abilities for critical thinking is an important educational goal. Critical-thinking educational goals focus on developing students who are fair-minded and objective, reach sound conclusions, and are disposed toward seeking clarity and accuracy (Marzano et al., 1992). What is critical thinking? Psychologists do not agree on all the skills that constitute it (Kuhn, 1999; Woolfolk, 2005). Discussions of critical thinking often use many of the same terms used in discussions of problem solving: The two areas are closely related.

In this chapter, we shall adopt the following definition: "**Critical thinking** is reasonable, reflective thinking that is focused on deciding what to believe or do" (Ennis, 1985, p. 54). This definition implies the following five attributes (Norris & Ennis, 1989):

1. *Reasonable thinking*—using good reasons
2. *Reflective thinking*—being conscious of looking for and using good reasons
3. *Focused thinking*—thinking for a particular purpose or goal
4. *Deciding what to believe or do*—evaluating both statements (what to believe) and actions (what to do)
5. *Abilities and dispositions*—both cognitive skills (abilities) and tendency to use the abilities (dispositions)

Critical-thinking abilities are specific cognitive skills that are used when a student exhibits critical-thinking behavior. Here are some of the abilities typically considered in discussions of critical thinking that could be assessed (Ennis, 1985). They are grouped into five areas.

FIGURE 11.6 **Strategies for assessing problem solving.**

Problem solving category	Strategy	Description
Identifying and recognizing problems	1. Identify the problem	Present a scenario or problem description. Ask students to identify the problem to be solved.
Defining and representing problems	2. Pose questions[1]	Present a statement that contains the problem and ask students to pose the question(s), using the language and concepts of the subject you are teaching, that need(s) to be answered to solve the problem.
	3. Demonstrate linguistic understanding	Present several problems students should be able to solve and underline the key phrases and common vocabulary they need to know to comprehend the context of the problem. Ask students to explain in their own words the meaning of these linguistic features of the problem.
	4. Identify irrelevancies	Present interpretive materials and a problem statement and ask students to identify all of the irrelevant information. Be sure the interpretive material contains information that is both relevant and irrelevant to the problem solution.
	5. Sort problem cards	Present a collection of two or more examples of each of several different types of problem statements and ask students to (a) sort the problems into categories or groups of their own choosing and (b) explain why the problems they put into a group belong together. Put each problem statement on a separate card, but do not specify the type of problem it is. Focus your assessment on whether students are attending to only the wording or other surface features of the problem or, more appropriately, to the deeper features of the problem. For example, students should group all problems that can be solved using the same mathematical principle, the same scientific law, etc., even though the problems are worded quite differently or are applied to different content.
	6. Identify assumptions	State a problem and ask students to state (a) a tentative solution and (b) what assumptions about the current and future problem situation they have made in reaching their solution.
	7. Describe multiple strategies	State a problem and ask students to (a) solve the problem in two or more ways and (b) show their solutions using pictures, diagrams, or graphs.
	8. Model the problem	State a problem and ask students to draw a diagram or picture showing the problem situation. Assess how the students represent the problem rather than on whether the problem is correctly solved. Drawings of time problems in mathematics, for example, should depict time lines, not scales.
	9. Identify obstacles	Present a difficult problem to solve, perhaps one missing a key piece of information, and ask students to explain (a) why it is difficult to complete the task, (b) what the obstacle(s) are, and/or (c) what additional information they need to overcome the obstacle(s). Assess whether students can identify the obstacle to solving the problem.
Exploring possible solution strategies	10. Justify solutions	Present a problem statement along with two or more possible solutions to the problem and ask students to (a) select one solution they believe is correct and (b) justify why it is correct.
	11. Justify strategies used	State a problem and two or more strategies for solving it, and ask students to explain why both strategies are correct. Be certain both strategies yield the correct solution. In writing an item you might, for example, state that these were different ways that two fictional students solved the problem.
	12. Integrate data	Present several types of interpretive material (story, cartoon, graph, data table) and a statement of a problem that requires using information from two or more of the interpretive material types. Then ask students to

Continued

FIGURE 11.6 (*Continued*)

Problem solving category	Strategy	Description
		(a) solve the problem and (b) explain the procedure they used to reach a solution. The problem solution must require using information from two or more of the interpretive materials.
	13. Produce alternate strategies	Present a problem statement and ask students to state two or more alternative solutions to the problem. An alternative approach is to present, along with the problem statement, one strategy that solves the problem, and require students to show you another way the problem could be solved.
	14. Use analogies	Present a problem statement and a correct solution strategy, and ask students to (a) describe other problems that could (by analogy) be solved by using this same solution strategy and (b) explain why the solution to the problem they generated is like the solution to the problem you gave them. Assess the analogical relationship of the students' solution strategy to the solution strategy you gave them.
	15. Solve backward	Present a complex problem situation or a complex (multistep) task to complete, and ask students to work backward from the desired outcome to develop a plan or a strategy for completing the task or solving the problem. For example, ask students to develop the steps and time frame needed to complete a library research paper. Assess how well students use backward solution strategies.
Acting on and looking back on problem-solution strategies	16. Evaluate the quality of a solution	State a problem and ask students to evaluate several different strategies for solving the problem. Ask students to produce several different solutions, or provide several solutions and ask them to evaluate those provided. If you provide solutions to evaluate, be certain to vary their correctness and quality, so that students can display their ability to evaluate. (For example, some may be more efficient, some may have negative consequences, and some may not work at all.) Ask students to determine the best strategy, explain why some strategies work better than others, and why some do not work at all. Assess the students' ability to justify the hierarchical ordering of the strategies' quality.
	17. Systematically evaluate strategies	Use the same types of tasks as in Strategy 16, but assess the extent to which students follow systematic procedures to evaluate each of the solution strategies you proposed.

[1] *Strategies 2, 7, 10, 11, and 12 were adapted from junior high school mathematics performance assessments described by Lane, Parke, and Moskal (1992). We stripped their definitions of mathematical content to suggest the general structure of the strategy.*

Elementary clarification

1. Focus on a question—Students who possess the ability to focus on a question can critically review an action, a verbal statement, a piece of discourse, a scientific or political argument, or even a cartoon to determine its main point(s) or the essence of the argument. Subskills include (a) formulating or identifying the question or issue being posed; (b) formulating or selecting the proper criteria to use in evaluating the material presented; and (c) keeping the issue and its proper context in mind.

2. Analyze arguments—Students who possess the ability to analyze arguments are able to analyze the *details* of the arguments presented in verbal statements, discussions, scientific or political reports, cartoons, and so on. The subskills include (a) identifying the conclusions in a statement; (b) identifying the stated and unstated reasons behind an argument; (c) seeing similarities and differences among two or more arguments; (d) finding, pointing out, and ignoring (when appropriate) irrelevancies appearing in an argument; (e) representing

the logic or structure of an argument; and (f) summarizing an argument.

3. Asking and answering questions that clarify and challenge—Students who possess the ability to ask clarifying questions can do two things: (a) ask appropriate questions of someone who is presenting an argument; and (b) answer critical questions appropriately when making an argument themselves.

Basic support

4. Judging the credibility of a source—Students with this ability can evaluate the quality of the evidence someone uses in supporting a position. Standards or criteria students should be able to use when judging credibility include (a) the expertise of the person giving the evidence; (b) whether the person giving the evidence has a conflict of interest; (c) whether different sources of evidence agree; (d) whether the source of evidence has a reputation for being accurate and correct; (e) whether the evidence was obtained by established procedures that give it validity; and (f) whether there are good reasons for using the evidence under the given circumstances. Each discipline will have specific rules of evidence, as well.

5. Making and judging observations—This is the ability of students to evaluate the quality of information obtained from eyewitness or direct observation of an event, phenomenon, or person. Among the standards or criteria students should be able to use when making these judgments are whether (a) an observer reports with minimal referral to others' observations; (b) the time between the event and the report by the observer is short; (c) an observer is not reporting hearsay; (d) an observer keeps records of the observation; (e) the observations reported are corroborated by others; (f) an observer had good access to the event or person so direct observation can be accurate; (g) an observer records the observations properly; and (h) an observer is a credible source.

Inference

6. Making and judging deductions—Students who are able to judge deductions apply logical thinking when they analyze statements and conclusions. Subskills include (a) using the logic of class inclusion (what elements or members should be logically included in a class or category); (b) using conditional logic (identifying the conditions under which something is true or false); and (c) properly interpreting statements using logical strategies (negatives; double negatives; necessary vs. sufficient conditions; and words such as *if, or, some, not, both*).

7. Making and judging inductions—Students who have the ability to induce can draw valid conclusions by generalizing from given information. Students who have the ability to judge inductions identify the conclusions that best explain the given evidence (Norris & Ennis, 1989). Subskills for generalizing from the data include (a) identifying and using typical features or patterns in the data to make inferences; (b) using appropriate techniques to make inferences from sample data; and (c) using patterns and trends shown in tables and graphs to make inferences.

8. Making and judging value judgments—Not all critical-thinking inferences are made using data and syllogisms. Some are based on judging value definitions. Students with this critical-thinking ability are able to identify when inferences have been made on the basis of values, what these values are, and when to use their own values to make inferences. Subskills of this ability include (a) gathering and using appropriate background information before judging; (b) identifying the consequences of the inferences that could be drawn and weighing the consequences before drawing conclusions; (c) identifying alternative actions and their value; and (d) balancing alternatives, weighing consequences, and deciding rationally.

Advanced clarification

9. Defining terms and judging definitions—Students who possess this ability are able to analyze the meanings and definitions of the terms used in the course of arguments, statements, and events to evaluate them critically. Among the subskills of this ability are (a) knowing the various forms that key terms may take and how these forms function in the context of an argument; (b) knowing how different strategies are used to define key terms in arguments; and (c) knowing the validity of the content of the definition itself.

10. Identifying assumptions—Students who possess this ability are able to identify assumptions that are part of someone's reasoning about what to believe or to do. In this case, we use the term *assumption* to mean an unstated basis for someone's reasoning. Be careful not to confuse this with

the common misconception that an assumption is a tentatively held conclusion.

Strategies and tactics

11. Deciding on an action—Students who can decide on an action are essentially good problem solvers. The subskills are those we discussed earlier in this chapter on problem solving: defining problems, formulating and evaluating solutions, viewing the total problem and taking action, and evaluating the action taken. The assessment strategies for this ability are the same as those you would use in assessing problem-solving skills.

12. Interacting with others—Students who are good at interacting with others are able to identify and use rhetorical devices to persuade, explain, or argue. Among the rhetorical devices students should be able to identify and use are (a) argumentative verbal tactics (appeal to authority, straw man, etc.); (b) logical strategies; and (c) skillful organization and presentation.

Source: From *Evaluating Critical Thinking* (p. 14), by S. P. Norris and R. H. Ennis, 1989, Pacific Grove, CA: Critical Thinking Books and Software. Reprinted by permission.

Strategies for Assessing Critical-Thinking Abilities

The ultimate goal of education in critical thinking is to enable students to use these abilities spontaneously in school and in their lives after school. For example, students would be expected to spontaneously clarify the main point of an argument that someone was stating unclearly by asking the person, "What is your main point?" or "Can I say that your main point is _______?"

For the most part, critical-thinking abilities are best taught and assessed in the context of individual subjects. For these reasons, and for the practical reason of limited space, we cannot illustrate meaningful items for assessing critical-thinking abilities in many different subjects. However, we do show the *strategies* you could use when crafting assessment tasks for these abilities. Some of these are illustrated with sample items. You need to practice applying these strategies to the subject(s) you teach.

The strategies shown in Figure 11.7 are organized around the headings used in the preceding list of critical-thinking abilities. Use these alone or in selective combinations that fit real-life circumstances. Give students instruction and practice in deciding on the appropriate analyses and when to use combinations of critical-thinking skills in different circumstances. Make the situations realistic. Performance tasks, discussed in Chapter 12, offer assessment opportunities for doing this.

Strategies for Assessing Dispositions Toward Critical Thinking

Dispositions toward critical thinking are habits of mind or tendencies to use appropriate critical-thinking behaviors often. Students who are disposed toward critical thinking:

1. seek a statement of the thesis or question;
2. seek reasons;
3. try to be well informed;
4. use credible sources and mention them;
5. take into account the total situation;
6. keep their thinking relevant to the main point;
7. keep in mind the original or most basic concern;
8. look for alternatives;
9. are open-minded and
   a. seriously consider points of view other than their own;
   b. reason from starting points with which they disagree without letting the disagreement interfere with their reasoning;
   c. withhold judgment when the evidence and reasons are insufficient;
10. take a position and change a position when the evidence and reasons are sufficient to do so;
11. seek as much precision as the subject permits;
12. deal in an orderly manner with the parts of a complex whole;
13. employ their critical thinking abilities;
14. are sensitive to the feelings, level of knowledge, and degree of sophistication of others.

Source: From *Evaluating Critical Thinking* (p. 12), by S. P. Norris and R. H. Ennis, 1989, Pacific Grove, CA: Critical Thinking Books and Software. Reprinted by permission.

Although you can assess students' use of a critical-thinking ability or skill on one occasion, *assessment of students' dispositions requires you to focus on their long-term habits*. Your assessment should report how frequently over a marking period, term, or year students use critical thinking in the curriculum subject matter. You assess dispositions using either a checklist or a rating scale.

FIGURE 11.7 **Strategies for assessing critical thinking.**

Critical thinking category	Strategy	Description
Elementary clarification	1. Focus on a question	Give students ■ Statement of a problem ■ Political address ■ Statement of a government policy ■ Experiment and results Ask students ■ What is the main issue/problem? ■ What criteria should you use to evaluate the quality, goodness, or truth of the argument or conclusions?
	2. Analyze arguments	Give students ■ Description of a situation *and* ■ One or two arguments Ask students ■ What conclusions are logically appropriate? ■ What evidence is presented that genuinely supports the argument(s)? ■ What evidence is presented that genuinely contradicts the argument(s)? ■ What are the unstated assumptions that need to hold for the argument(s) to be valid? ■ What part(s) of the statement is irrelevant to the argument(s)? ■ Outline the logical structure of the argument(s). ■ Summarize the main parts of the argument(s).
	3. Ask clarifying questions	Give students ■ Description of a situation *and* ■ An argument Ask students ■ What question(s) would you ask of the speaker or author? ■ Why would you ask these things?
Basic support of an argument	4. Judge the credibility of a source	Give students ■ Texts of arguments ■ Advertisements ■ Experiments and interpretations Ask students ■ Which parts, if any, of the material are credible, and why? ■ Which parts of the material are not credible, and why?
	5. Judge observation reports	Give students ■ Description of the context for the observation *and* ■ Report(s) of the observation(s) *and* ■ Background of the observer or reporter Ask students ■ Can you trust or believe the report of _____? ■ Which parts of the material are not credible, and which are not? Why?
Inferences	6. Judge deductions by a. comparing different conclusions b. judging the truth of a conclusion	(a) Give students ■ A statement that students are to assume is true *and* ■ Alternatives consisting of one logically correct conclusion and two or more logically incorrect conclusions Ask students ■ Which conclusions follows?

Continued

FIGURE 11.7 (*Continued*)

Critical thinking category	Strategy	Description
		(b) Give students ■ A statement that students are to assume is true *and* ■ An alternative consisting of one possibly correct conclusion Ask students ■ Must this conclusion follow?
	7. Judge inductions	(a) *For response-choice items,* give students ■ Situation statement *and* ■ Information (data) *and* ■ Possible conclusions drawn from the information Ask students ■ Judge the conclusion as supported or contradicted (or neither) by the data *or* ■ Select the conclusion that best explains the data. (b) *For constructed-response items,* give students ■ Situation statement *and* ■ Information (data) Ask students ■ Draw the proper conclusion from the data *and* ■ Explain why the conclusion is correct
	8. Make judgments about values	Give students ■ Description of a situation *and* ■ Problem statement *and* ■ Possible solutions to the problem Ask students ■ What are the positive and negative consequences of each solution? ■ Which is the most valuable solution, and why?
Advanced clarification	9. Judge definitions	Give students ■ Situation statement *and* ■ Argument or discourse Ask students ■ Analyze the way the speaker uses key terms to affect the listener ■ Explain how the definitions of the key terms are used in the argument to convince the listener ■ In the selection, identify all shifts of meaning of the key term ______. What effect does the shift have? Why did the speaker shift meanings?
	10. Identify implicit assumptions	Give students ■ An argument or explanation with some of its bases not included *and* ■ One option that is the correct implicit assumption *and* ■ Two or more options that are not the implicit assumption *and* that are not conclusions Ask students ■ Which option is probably assumed? ■ Which option is probably taken for granted?
Strategies and Tactics	11. Decide on an action	[This is essentially problem solving. Use the strategies for assessing problem solving in Figure 11.7.]
	12. Interact with others	Use performance assessment (e.g. a debate) and a scoring scheme, most likely a rubric or rating scale.

FIGURE 11.7 **(*Continued*)**

Critical thinking category	Strategy	Description
	13. Identify rhetorical mechanisms and tactics	Give students ■ Persuasive writing, a speech, an advertisement ■ A video clip of a speech or advertisement Ask students ■ What deceptive or misleading statements or strategies are used? Explain. ■ Which of the following types of deceptive or misleading statements or strategies are used?

Source: Outline is from *Evaluating Critical Thinking* (Table 1.2, p. 14), by Stephen P. Norris and Robert H. Ennis, 1989, Seaside, CA: Critical Thinking Co. 800-458-4849/www.criticalthinking.com. Reprinted by permission.

Checklists A **checklist** is a tool that contains a list of behaviors. You observe students over a period of time and make a checkmark (✓) next to the behavior you have observed. You then have a record of which disposition behaviors students have exhibited. The more behaviors you checked, the greater the students' dispositions toward critical thinking. Chapter 12 gives specific suggestions for crafting these types of assessment devices. An example checklist that could be used to assess a student's dispositions toward critical thinking is shown in Figure 11.8.

This checklist could help you keep track of a student's critical-thinking actions over the course of a unit. You can see from the checklist that the student exhibited a number of dispositions frequently (e.g., "2. Looks for explanations and reasons," "6. Open-minded") and others not very frequently (e.g., "5. Looks for alternatives"). You can use this information to help the student develop his critical thinking by teaching him to develop the habit of always looking for alternatives.

Rating Scales A simple **rating scale** is a device to record your judgments of the quality level of a student's dispositions toward each critical-thinking behavior. A rating scale usually has a line with points on it that range from poor quality to excellent quality. Usually, four or five quality points are further defined by describing what the behavior looks like at each point. These descriptions are called *anchors*. Figure 11.9 is an example of some of the dispositions toward critical thinking that a teacher might observe as a student completes an assignment. The anchor points on the items' rating scales were adapted from the Marzano, Pickering, and McTighe (1993) analysis of habits of mind.

In this example, each item's scale shows the degree to which a student is disposed toward using a particular critical-thinking habit. The numerical ratings on the scale are anchored by descriptions of specific and observable behaviors. Over time you can observe the student with respect to these habits. Then, at the end of the period, you use the rating scale to assess the student's disposition on each habit.

Finally, if students are to learn to be disposed toward using critical thinking in their daily activities, you should teach students critical-thinking skills. Assess both critical-thinking skills and dispositions continuously throughout the term or year.

READING SKILLS

Reading skills involve thinking, too. Three strategies for assessing reading skills are presented.

Traditional Procedure Having students read material in a subject area and answer questions based on that material is a desirable way to assess reading skills. Although developing passages followed by questions is not easy, you may need to do so, especially when teaching subjects for which you lack adequate assessment procedures or study booklets covering these skills. To develop such assessments, the reading materials need to be carefully selected to represent the kind of material students should be able to read. Also, the reading material may need to be rewritten so that the interpretive questions can be answered primarily on the basis of

FIGURE 11.8 **Example of a checklist to keep track of a student's use of critical-thinking dispositions throughout a teaching unit.**

Individual Student's Critical-Thinking Disposition Record

Student's name: **Class period:** **Dates:**

Subject/unit: U.S. History/Unit III. Beginning a Government, 1780–1800

	Assignment/activity				
Critical-thinking dispositions	Class discussion of the Articles of the Confederation	Essay discussing arguments for and against ratification of the Constitution	Scrapbook collecting and analyzing events reported in the newspaper using concepts from the Constitution	Teams debate the issue, "Have political parties made the United States goverment better?"	Essay evaluating Washington as president
1. Seeks statements of the main point or question	✓	—	✓	✓	NA
2. Looks for explanations and reasons	✓	✓	✓	✓	✓
3. Uses and cites credible sources	—	✓	—	✓	—
4. Keeps to the main and relevant point(s)	—	—	NA	✓	✓
5. Looks for alternatives	—	—	NA	—	NA
6. Open-minded	✓	✓	✓	NA	—
7. Takes a position on an issue		✓	—	✓	—
8. Changes position on an issue with good reason(s)	NA	—	NA	NA	NA
9. Seeks to be accurate and precise in statements and work	NA	—	✓	✓	—
10. Sensitive to the feelings, levels of knowledge of others	✓	NA	NA	—	NA

FIGURE 11.9 Sample rating scale assessing the quality of some of a student's dispositions toward critical thinking that a teacher might observe as the student completes an assignment

Rating Scale for Critical Thinking Dispositions

Student's name: Date:
Assignment:

1. Did the student consider different points of view?

0	1	2	3
Acts as if own point of view is accepted by everyone	Aware that own point of view is not accepted by everyone	Shows awareness that others have legitimate points of view that differ from own	Actively looks for and encourages others to express points of view different from or opposing own

2. How did the student treat others' points of view?

0	1	2	3
Acts in a way that avoids or discourages others' points of view	Shows some attention to others' points of view	Makes a serious effort to consider others' points of view, but is not consistently objective or rational	Attends seriously to others' points of view and consistently reviews them objectively and rationally

3. Did the student communicate well with others who had less knowledge or ability?

0	1	2	3
Cannot work or communicate with others who have less knowledge or ability	Attempts to work or communicate with others who have less knowledge or ability, but is less than adequate in doing so	Works and communicates adequately with others who have less knowledge or ability	Works and communicates excellently with others who have less knowledge or ability

4. Was the student sensitive to the feelings of others with less knowledge or ability?

0	1	2	3
Acts apathetically or cruelly toward others who have less knowledge and ability	Does the minimum to help or encourage respect for the feelings of others who have less knowledge and ability	Offers good encouragement and respect for the feelings of others who have less knowledge and ability	Actively seeks to bolster and increase respect for the feelings of others who have less knowledge and ability

the material presented. Finally, questions need to be phrased in a way that does not require a student to have more background or special information than you deem appropriate for the level of students and subject matter at hand (Wesman, 1971).

The steps for building a set of assessment questions requiring reading and interpreting of a printed passage follow (Wesman, 1971):

1. *Locate a promising passage.* Examine sources (texts, periodicals, reference works, specialized books, and collections and anthologies) until you find a passage for which you can write several interpretive items.
2. *Write initial test items.* Write as many items for the passage as you can. Try to exploit all of the possibilities for interpretation of the passage that fit your original assessment plan.
3. *Rewrite the passage.* After you have a tentative set of items, rewrite the passage to eliminate unessential material that does not contribute to the items you have written.
4. *Consider rewriting some of the items.* Changes in the passage may require revising or eliminating some of the items you already wrote. The goal of Steps 3 and 4 is to produce a condensed and efficient passage and item set.
5. *Repeat Steps 3 and 4 as often as necessary,* until you are satisfied that you have an efficient set of items.

Most commercial survey achievement tests contain reading comprehension subtests. You should consult these for examples of using passages to assess reading comprehension.

Authentic Reading Assessment If you are most interested in reading comprehension, rather than the students' ability to interpret subject-matter materials, the preceding type of condensation may be undesirable, especially if part of what you want to assess is the ability to read naturally occurring materials and the capacity to distinguish between relevant and extraneous material (Wesman, 1971). Critics of standardized reading comprehension tests argue that the passages and questions created by the traditional 5-step procedure are too artificial. The critics would rather use materials that students need to read in the real world or in further schooling. They claim that students exposed to traditional reading comprehension tests come to believe that reading (a) consists of short passages, (b) requires answering questions whose answers are known by the authorities that set them, and (c) has little to do with interpreting the written word (Resnick, 1989).

Passages are considered authentic if they are drawn from the primary sources of a discipline, age-appropriate books and magazines, newspapers, and textbooks students may encounter. In addition, authentic reading tasks may require students to read longer passages than typically appear on traditional reading comprehension tests. They may also require students to read from several sources to compare points of view or obtain reliable and complete information. For example, a student may read four different accounts of an event or of a procedure and then answer questions about the event or procedure, or about comparisons among the different accounts read.

Alternatively, you may want to combine reading, writing, and subject-matter exercises, such as conducting science experiments. For instance, students can individually read several pieces and answer questions about them. Then you can organize students into groups to discuss the pieces they read and to share their insights into interpreting them. Next, students can individually write essays or set up experiments to extend or synthesize the material they have read and discussed. The purpose of the intermediate discussion is to offer all students the opportunity, through group discussion, to clarify points, obtain information they may have missed through their reading, and "level the playing field" somewhat before the writing phase of the assessment begins.

Longer and more authentic reading tasks use more of your class time for assessment than does the traditional method. An assessment that requires reading several original texts and writing essays after a class discussion may take several class periods to complete. You need to balance your assessment time against your teaching time before deciding which assessment strategy to use. You could try some combination of both. You could use more authentic assessment methods on some occasions and more efficient assessment methods on others. Compare students' performance and the type of information you obtain under the different approaches. This may give you some insight into the validity of the assessment results from each method.

The MAZE Item Type Reading comprehension can be assessed through a multiple-choice variety of the **cloze reading exercise** known as the **MAZE item type**. The basic idea is to find an appropriate passage and embed a multiple-choice question in the passage that students can answer only if they comprehend the meaning of the surrounding passage. To better understand this procedure, consider the following multiple-choice item, which requires students to select the word that best completes the sentence, "The baby ________."

Example

MAZE multiple-choice item before it is embedded in text

1. The baby _____.
   A cried
   B laughed
   C slept
   D walked

Notice that all options correctly complete this sentence when it is read outside the context of a reading passage. Now, consider the same item when it is embedded in a brief passage as shown in Item 2 below:

Example

MAZE multiple-choice item after it is embedded in text

2. Mother and her six-month-old baby played for a long time. The baby _____. He enjoyed being tickled under the arms.
   A cried
   *B laughed
   C slept
   D walked

Option B is correct because of the context in which the item is embedded. Item 2 shows a simple paragraph of a few short sentences; this technique also can, and should, be applied to longer and more complex prose passages.

MAZE items appear to have a considerable advantage over the usual cloze exercises, in which only the blank appears and students must fill in the missing word. They (a) assess whether students can construct meaning from the passages, (b) are objectively scored, (c) do not result in a student filling in blanks with words that leave you wondering whether a student understands the passage, and (d) do not require students to have a great deal of outside knowledge for you to assess their ability to read.

The following suggestions for formulating MAZE test items are based on those used for the *Degrees of Reading Power* (Touchstone Applied Science Associates, 1995a, 1995b) test.

1. *Design the items so that a student needs to read and understand the passage to answer correctly.* As in the preceding example, when an item is considered in isolation, each option should make the sentence grammatically and semantically correct. However, once the item is embedded in the text, only one option should be correct.
2. *The passage should contain all the content information a student needs to answer the item correctly.* For the item to assess reading ability, a student should not have to depend on recall of special experiences to find the correct answer. This usually means the passages must be written specifically for the test.
3. *All of the items' options should be common words.* All students should recognize and understand the meaning of each option. This ensures that when a student misses an item, the fault lies with the student's inability to comprehend the reading passage rather than the student's lack of knowledge about the meaning of the words in the item.

CONCLUSION

We have discussed strategies for assessing four aspects of higher-order thinking skills: concept learning, rule-governed or principled thinking, problem solving, and critical thinking. These categories overlap. Our separate presentation of each aspect was crafted to explain assessment strategies, not to imply there are four separate "kinds" of thinking. However, we believe this toolkit-style list of assessment strategies will come in quite handy. In the next chapter, we turn to performance assessments. These, too, can be used to assess higher-order thinking. They also assess students' abilities to create and construct products.

EXERCISES

1. Identify a principle or rule in a subject you teach or plan to teach. Then complete these tasks:
   a. State the subject and the grade level.
   b. State the principle and give its name.
   c. Describe in general terms the conditions under which it is appropriate to use the principle to solve a problem or explain a phenomenon.
   d. Using general terms, describe the most likely kinds of faulty inferences made or conclusions drawn by students who misinterpret or misapply the principle.
   e. Prepare one multiple-choice item to assess a student's ability to identify an appropriate conclusion to be made when applying this rule. Use the information in your answer to the previous question as a basis for formulating distractors. Use the checklist for judging the quality of multiple-choice items to improve your item.
   f. Prepare a constructed-response item to assess a student's ability to produce examples of conclusions after applying the principle. This item should assess the same principle that you used for constructing the multiple-choice item just written.
   g. Administer both of these items, one at a time, to a student at the appropriate grade level. Administer the constructed-response item first, then remove it from the student before administering the multiple-choice item.
   h. Compare the results you obtained. What were the similarities and differences in the quality of information you received? Which task is more valid? Why?
   i. Share your results with others in this class. How do your results compare with theirs? Were there differences with respect to subject matter and grade level assessed? What conclusions can the class draw from its collective experience?
2. For the subject you teach or plan to teach, develop a notebook with well-designed tasks that assess different problem-solving abilities. Organize your tasks according to the categories of the IDEAL problem solver. Structure your notebook as follows:
   a. Create one assessment task using each of the 17 assessment strategies for problem-solving assessment presented in Figure 11.7.
   b. Type one assessment task per page. Label the task with the subject, teaching unit, student

grade level, assessment strategy, and category of the IDEAL problem solver the task assesses.

c. On a separate page, craft a scoring rubric for the task and write a sample ideal response.

d. Self-assess your work. Be sure that the content of each task and the scoring rubric are accurate and that the tasks are well crafted.

e. Share your notebook with the other participants in this course.

3. Select a subject and grade level you teach or plan to teach, for which critical thinking is an important learning outcome. Then, complete the following tasks:

a. Identify and briefly describe (in general terms) one or more teaching units in which critical-thinking abilities can be taught and practiced.

b. On a large sheet of paper, create a table in which each row heading is one of the 13 critical-thinking abilities listed in Figure 11.7.

c. Label the columns with the teaching and learning activities in the unit(s) that lend themselves to teaching and practicing critical-thinking abilities.

d. For each cell in the body of the table, briefly describe how a student would demonstrate that he or she was engaging in the corresponding critical-thinking ability. Not every cell will be filled, because not every ability can be demonstrated with every activity you list. However, in the table as a whole, all abilities should be demonstrated at least once. If they are not, then add a unit or an activity to your table.

e. Present your table to the others in this course. Discuss your activities and demonstrations. Revise your table on the basis of the discussion. Then share it with the other class members.

4. For the same subject and grade level you identified in Exercise 3, develop a notebook containing samples of tasks assessing each of the critical-thinking abilities listed in Figure 11.7. Structure your notebook around the 13 abilities as follows:

a. Craft one assessment task using each of the critical-thinking assessment strategies described in this chapter.

b. Type one assessment task per page. Label the task with the subject, teaching unit, student grade level, assessment strategy, and critical-thinking ability the task assesses.

c. On a separate page, craft a scoring rubric or scoring guide for the task and write a sample ideal response.

d. Review your work carefully. Be sure the content of each task and scoring rubric is accurate. Be sure the tasks are well crafted.

e. Share your notebook with the other members of this course.

5. For a subject and grade level you teach or plan to teach, identify a graph and a table (chart) the students should be able to use:

a. Craft a context-dependent item set for the graph assessing the students' ability to use it beyond simply reading values from it. Craft at least two tasks for the set. Review the items using the appropriate checklists from Chapters 8 and 9. Attach a completed checklist for each item.

b. Share your context-dependent item sets with the other members of this course.

Performance and Portfolio Assessments

KEY CONCEPTS

1. A performance assessment (a) requires students to create a product or demonstrate a process, or both, and (b) uses clearly defined criteria to evaluate the qualities of student work.
2. Types of performance tasks range from structured, on-demand tasks to longer-term projects and portfolios.
3. Advantages of performance assessment stem from its ability to assess complex learning targets. Disadvantages of performance assessment stem from the difficulties arising from that complexity.
4. To create a performance assessment, first be clear about the performance you want to assess.
5. The second step in creating a performance assessment is to design the task to elicit the desired performance.
6. The third step in creating a performance assessment is to design a scoring scheme that reflects the performance criteria. The scoring scheme may be one of several types of rubrics, a checklist, or a rating scale.
7. Projects can be used as performance assessments if they are designed according to the three steps (clearly describe the performance to be assessed, then match both the project task and scoring scheme to it).
8. For purposes of assessment, a portfolio is a limited collection of a student's work used either to present the student's best work(s) or to demonstrate the student's educational growth over a given period of time.

IMPORTANT TERMS

adaptive assessment task
alternate solution strategies
alternative assessment
authentic assessment
behavior checklist
central tendency error
combined group and individual project
debate
demonstration
descriptive graphic rating scale
dramatization
electronic portfolio
exemplars
experiment
graphic rating scale
group project
halo effect
naturally occurring performance
leniency error
logical error
non-paper-and-pencil task
numerical rating scale

From Chapter 12 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

on-demand task

oral presentation

paper-and-pencil task

performance task

personal bias

portfolio

portfolio culture model

procedure checklist

product checklist

reliability decay

reliability of ratings

rubrics (analytic, annotated holistic, general (generic), holistic, task-specific)

scaffolding

scenario

scoring rubric

self-evaluation checklist

severity error

simulation

standardized patient format

structure a task

structured task (exercise)

WHAT IS PERFORMANCE ASSESSMENT?

A performance assessment (a) requires students to create a product or demonstrate a process, or both, and (b) uses clearly defined criteria to evaluate the qualities of student work. A performance assessment requires students to do something with their knowledge, such as make something (build a bookshelf), produce a report (report on a group project that surveyed parents' attitudes), or demonstrate a process (show how to measure mass on a laboratory scale). Figure 12.1 shows a decision-making task that a history teacher constructed.

A performance assessment must have two components: the performance task itself and a clear rubric for scoring. The rubric should be based on stated learning targets. Classroom instructional activities that lack this scoring rubric component do not qualify as performance assessments. In practice, this line can become blurred; as good teachers observe students working, they sometimes do have criteria in mind and use their observations for formative purposes. When you teach, you use many learning activities to engage students' interest, to give them experience, and to give practice with the learning targets. For valid information from summative, graded performance assessments, both task and criteria are necessary.

The Performance Task

A **performance task** is an assessment activity that requires students to demonstrate their achievement by producing an extended written or spoken answer, by engaging in group or individual activities, or by creating a specific product. When you use a performance task, you require students to demonstrate directly their achievement of a learning target. You require only indirect demonstration if you ask students simply for a brief answer (e.g., completion items or short-answer) or to select an answer from among options you present to them (matching exercises, true-false items, or multiple-choice items).

To choose an assessment method, you must first be clear what learning target you want to assess, then match the method to it. Learning targets that require students to learn facts (such as the structure of the nation's government or the main provisions of the Constitution and the Bill of Rights), comprehend an event or a theory, or compare two or more concepts lend themselves well to traditional item formats. Use performance tasks, on the other hand, to assess learning targets that require students to apply their knowledge and skills as they perform something. A good rule of thumb to remember is that simple learning targets require simple assessment formats; complex learning targets require complex assessment (Arter, 1998). Use many different types of assessment procedures (short-answer items, objective items, and a variety of long-term and short-term performance tasks) to sample the breadth of your state's standards and your local curriculum's learning targets.

Performance assessment is sometimes called **alternative assessment** or **authentic assessment**. These terms are not interchangeable, however. The "alternative" in *alternative assessment* usually means in opposition to standardized achievement tests and to multiple-choice (true-false, matching, completion) item formats. The "authentic" in *authentic assessment* usually means presenting students with tasks that are directly meaningful to their education

FIGURE 12.1 **A performance assessment task for a high school history course.**

DECISION-MAKING TASK

Suppose you lived in the 1940s and President Harry S Truman requested that you serve on a White House task force. The goal is to decide on how to force the unconditional surrender of Japan, yet provide for a secure postwar world.

You are a member of the committee of four and have reached the point at which you are trying to decide whether to drop the atomic bomb. Identify the alternatives you are considering and the criteria you are using to make the decision. Explain the values that influenced the selection of the criteria you are using to make the decision. Also explain how your decision has helped you better understand this statement: “War forces people to confront inherent conflicts of values.”

Before you begin your task, establish a clear goal and write it down. Then write down a plan for accomplishing your goal. When you are finished with the task, be prepared to describe the changes you had to make in your plan along the way.

As you work on your task, try a variety of sources of information: books, magazine articles, newspapers, and people who lived through the war. Keep a list of those sources and be prepared to describe how you determined which information was most relevant and which information was not very useful. Present your conclusions and findings in at least two of the following ways:

- A written report
- A letter to the president following the completion of the committee meeting
- An article written for *Time* magazine, complete with suggested photos and charts
- A videotape of a dramatization of the committee meeting
- An audio tape
- A newscast
- A mock interview

You will be provided rubrics for and be assessed on each of the following learning targets:

Content Learning Target

1. Your understanding of the principle that war forces sensitive issues to surface and causes people to confront inherent conflicts of values and beliefs.

Information Processing Learning Targets

2. Your ability to review and evaluate how valuable each source of information is to the parts of your project.

Complex Thinking Learning Targets: Decision Making

3. Your ability to identify important and appropriate alternatives to dropping the bomb.
4. Your ability to identify important and appropriate criteria to evaluate each alternative to be considered.
5. Your ability to accurately evaluate the extent to which each of your alternatives meets each criterion.
6. Your ability to select the alternative(s) that adequately meets (meet) the criteria and answer the initial decision question.

Habits of Mind Learning Target

7. Your ability to effectively define your goal in this assignment and to explain your plan for attaining this goal.

Effective Communication Learning Target

8. Your ability to communicate your conclusions and findings in two or more ways.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (pp. 27–29), by R. J. Marzano, D. Pickering, and J. McTighe, 1993, Alexandria, VA: Association for Supervision and Curriculum Development. Adapted by permission of McREL, 4601 DTC Blvd. #500, Denver, CO 80237.

instead of indirectly meaningful. For example, reading several long works and using them to compare and contrast different social viewpoints is directly meaningful because it is the kind of thoughtful reading educated citizens do. Reading short paragraphs and answering questions about the “main idea” or about what the characters in the passage did, on the other hand, is indirectly meaningful because it is only one fragment or component of the ultimate learning target of realistic reading. “Realistic” and “meaningful” are terms educators writing about authentic assessment often use. They beg some questions: “Realistic in which context?” and “Meaningful for whom?” And of course, there are degrees of authenticity; it is not an all-or-nothing concept.

Be sure that what your performance assessment requires students to do matches the learning targets—including appropriate thinking targets—and that your scoring rubrics evaluate those same learning targets. For example, if the learning target says that students must weigh chemicals on the laboratory scale, the performance task must require actual weighing, not an essay on how to use the scale. In addition, you must have a scoring rubric to help you decide how well students weigh the chemicals and not simply specify that chemicals were weighed.

The Rubrics for Scoring

A **scoring rubric** is a coherent set of rules you use to assess the quality of student performance: The rules guide your judgments and ensure that you apply your judgments consistently. The rules may be in the form of a rating scale or a checklist. Complex performances require that you assess several learning targets or several parts of the performance. To do this, you use several scoring rubrics: one for each learning target or part.

A *rating scale* consists of numerals, such as 0 to 3, or 1 to 4, that reflect the quality levels of performance. Each numeral corresponds to a verbal description of the quality level it represents. (We present an example in the following paragraphs.) Instead of verbal descriptions, examples of students' work may serve as concrete illustrations of each quality level.

TYPES OF PERFORMANCE ASSESSMENTS

Many types of tasks fit the broad definition we adopted here. Figure 12.2 lists most of these and gives examples.

Structured, On-Demand Tasks

For **structured, on-demand tasks** or exercises, the teacher specifies the task and materials for performance, describes the kinds of outcomes toward which students should work, tells the students they are being assessed, and gives students opportunities to prepare themselves for the assessment. Such tasks are also called on-demand (or controlled) tasks.

Paper-and-Pencil Tasks

We have already studied many types of **paper-and-pencil tasks**. In Chapters 10 and 11 we discussed constructed-response and essay items. These formats permit students not only to record their answers but also to give explanations, articulate their reasoning, and express their own approaches toward solving a problem. Sometimes your main focus is on the written product itself, such as the stories, reports, or drawings students create. At other times, you may be more interested in the process students use, for example, when students record the steps they used to complete an experiment or explain how they solved a problem.

Tasks Requiring Other Equipment and Resources

In subjects such as mathematics, science, mechanical drawing, art, first aid and life-saving, consumer science, and driver's education, important outcomes require students to do something with equipment and resources rather than write about how to do it. In some academic subjects, performing a **non-paper-and-pencil task** might be a better option than using a written response, even though either could be done.

For example, in elementary school general science, you would *directly assess* students' understanding and use of the metric system if you required them to measure objects, volume, mass, and so on. You use *indirect assessment* if you require students only to perform numerical conversions from one system or unit of measurement to another, or to answer questions based on pictures of measuring equipment. You could assess students' estimation skills, measuring skills, and systematic thinking skills, for example, all in one task by giving students a jar of beans and some simple tools. After giving students suitable directions, you can observe how they solve the problem of estimating the number of beans in the jar.

You may use non-paper-and-pencil tasks to present problems to be solved by a group, an individual, or some combination of group and individual work. In the latter case, the group may work cooperatively on the task; after the group solves the problem, individuals describe or write up what the group did and the solution to the problem. Non-paper-and-pencil tasks may also be open response, allowing for alternative correct performances, or closed response, allowing only one best or correct answer.

Demonstrations

A **demonstration** is an on-demand performance in which a student shows he can use knowledge and

FIGURE 12.2 **Common types of performance assessment techniques and examples.**

Type of performance assessment		Example*
Structured, on-demand tasks for individual students, groups, or both The teacher decides what and when materials should be used, specifies the instructions for performance, describes the kinds of outcomes toward which students should work, tells the students they are being assessed, and gives students opportunities to prepare themselves for the assessment.	Paper-and-pencil tasks	■ Solve this arithmetic story problem and explain how you solved it. ■ Study the following graph that shows how Sally uses her time. Then, write a story about a typical day in Sally's life using the information from the graph. ■ Draw a diagram to illustrate the mathematical ideas in the following word problem.
	Tasks requiring equipment and resources beyond paper and pencil	■ Build as many geometric shapes as possible from this set of four triangles. ■ Talk on this telephone to ask about a job and to request a job application. ■ Show me how to mix acid and water.
	Demonstrations	■ Demonstrate the proper way to knead dough for bread. ■ Demonstrate how to set up the microscope for viewing stained slides. ■ Demonstrate how to climb a rope. ■ Demonstrate how to look up information on the Internet.
Naturally occurring or typical performance tasks Observe students in natural settings: in typical classroom settings, on the playground, or at home. In natural settings you have to wait for the opportunity to arise for a particular student to perform the particular activity you would like to assess. The activity may not occur while you are observing.		■ Observe a student's way of dealing with conflicts on the playground. ■ Collect all pieces that each student wrote in every subject and analyze them for grammatical, spelling, and syntactic errors to determine a student's typical language usage (at least in school assignments). ■ Observe whether a student makes change correctly when running a refreshment stand at the school fair.
Longer-term projects for individual students, groups, or both You can combine group and individual projects. Groups of students can work on a long-term project together; after the group activities are completed, individuals can prepare their own reports. The combination approach is useful when a project is complex and requires collaboration to complete in a reasonable time frame, yet the learning targets require individual abilities.	Long-term reports	■ Collect and classify newspaper and magazine advertisements in the months before each holiday during the semester. ■ Using resources in the school library, write a research paper on why voter turnout is so low during primary elections. ■ Write a term paper on everyday life in Colonial America
	Tasks requiring equipment and resources beyond paper and pencil	■ Make a diorama depicting everyday life in Colonial America ■ Build a model of the solar system ■ Build a small piece of furniture using the hand tools you learned to use during the semester. ■ Build a working model of a camera using the optical principles taught in this unit.
	Experiments	■ Plan, conduct, and report on a study to answer the question, "Do most students in this school support the death penalty laws in this state?" ■ Plan, conduct, and report on an experiment to investigate the hypothesis that a brightly colored advertisement will be remembered longer than a dull one.
	Oral presentations and dramatizations	Write and present a skit depicting everyday life in Colonial America
Simulations	Actors and "standard patients" Computerized adaptive audio-visual scenarios Computerized adaptive text scenarios Computerized audio-visual simulations	Diagnose pathology in certain organs and organ systems using anatomy simulation software, and describe conventional treatment(s).

*Examples are general descriptions for the purpose of illustration. Actual performance assessments for students would need clear directions, more specifics about the task, and scoring criteria.

skills to complete a well-defined, complex task. Demonstrations are not as long or as complex as projects. Demonstrations are usually closed-response tasks. Tasks comprising a demonstration are often well defined and the "right" or "best way" is often known to both the student and the evaluator. However, individual variations are permitted; style and manner of presentation often count when a student presents a demonstration. The 4-H Clubs often use demonstrations: Boys and girls demonstrate their skills in a variety of agricultural and homemaking areas.

For the most part, demonstrations focus on how well a student uses her skills, rather than on how well the student can explain her thinking or articulate the principles underlying a phenomenon. If you use a demonstration for assessment purposes, you should carefully identify the appropriate learning target and use an appropriate scoring rubric.

Naturally Occurring or Typical Performance Tasks

In opposition to structured, on-demand performances are naturally occurring tasks. **Naturally occurring performances** require you to observe and assess students in natural settings: in typical classroom settings, while on the playground, or while at home. In these settings you are likely to see the way a student typically performs on a learning target, such as cooperating with members of a group to achieve a goal. In natural settings you do not tell students they are being assessed, nor do you control the situation in any way.

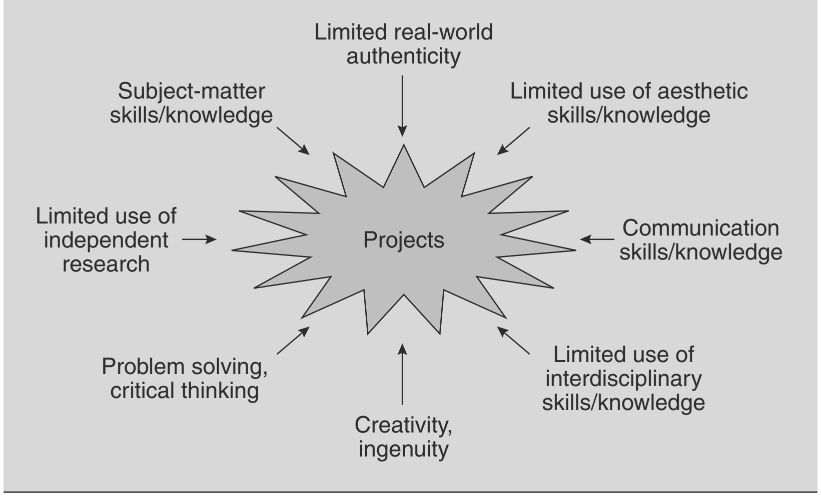
Although a naturally occurring setting may let you assess students' typical performance, this is not always the case. In natural settings you often have to wait for the opportunity to arise for a particular student to perform the particular activity you would like to assess. The activity may not occur while you are observing. This waiting lowers the efficiency of this assessment mode. For example, collecting writing assignments is unlikely to provide you with all the information you need to determine your students' command of the mechanics of writing. Not all spelling patterns and forms of sentence structure students need to learn, for example, are likely to appear in every student's writings. Thus, you would have no way of thoroughly assessing students' use of sentence structures. Formal assessment of performance learning targets usually requires a structured performance assessment.

Longer-Term Projects

Individual Student Projects An **individual project** is a long-term activity that results in a student product: a model, a functional object, a substantial report, or a collection.

Figure 12.3 shows that properly crafted projects require students to apply and integrate a wide range of abilities and knowledge; and use creativity, originality, and some sense of aesthetics. When students write a library research paper, for example, they must apply the skills of locating and using reference materials and sources: outlining, organizing, and planning a report; communicating using written language, word processing, and presentation style; and demonstrating their understanding

FIGURE 12.3 **Features of projects.**


Limited real-world authenticity
Subject-matter skills/knowledge
Limited use of aesthetic skills/knowledge
Limited use of independent research
Projects
Communication skills/knowledge
Problem solving, critical thinking
Limited use of interdisciplinary skills/knowledge
Creativity, ingenuity

of the topic. A good project will engage students in critical thinking, creative thinking, and problem solving.

Group Projects A **group project** requires two or more students to work together on a longer project. The major purpose of a group project *as an assessment technique* is to evaluate whether students can work together cooperatively and appropriately to create a high-quality product. The learning targets for a group project depend on the subject matter and the level of the students you are assessing. For example, group projects may focus on:

- *Action-oriented learning targets* (creating a newsletter)
- *Student-interest-oriented learning targets* (writing a paper on a topic they're interested in)
- *Subject-matter-oriented learning targets* (understanding how rivers are formed)
- *Interdisciplinary learning targets* (designing a wildlife refuge)

An example of a subject-oriented group project is shown in Figure 12.4.

To assess students' group learning skills, develop scoring rubrics. Rubrics are discussed later in this chapter, but for the sake of an example here, Figure 12.5 displays a set of general scoring rubrics for assessing collaboration and cooperation during group tasks. These rubrics could be adapted to a specific project.

Combining Group and Individual Projects In a **combined group and individual project**, groups of students work on a long-term project together, and

FIGURE 12.4 **A group project in a U.S. history course for students in middle school or high school.**

HISTORICAL INVESTIGATION TASK

In recent years controversy has arisen over the status of Christopher Columbus. Was he a hero or villain? As we study Columbus we will read from a number of resources penned by different historians that will present their views of Columbus.

In cooperative groups, choose at least two resources that describe conflicting reports of events that took place upon Columbus's "discovery" of the New World and during its settlement. Discuss the contradictions you find and try to determine why the historians reported the events differently. Using the resources available, develop a clear explanation of the reasons for the contradictions or present a scenario that clears up the contradictions.

Your group will explain to the class why historians seem to report the same event differently. In addition, your group will offer to the class its ideas for resolving the contradictions. Your group's presentation to the class may be either a dramatization, a panel discussion, or a debate.

Your project will be due 3 weeks from today. Every Friday one member of your group will tell the class the progress you made on the project during the past week, any problems the group had in completing the assignment, and what the group plans to complete during the next week.

Each member of the group will be assessed on the learning targets that follow. You will be provided rubrics for each of the learning targets so you may see more clearly what the assessment will be.

Social Studies Content Learning Target

1. Your understanding that recorded history is influenced by the perspective of the historian.
2. Your understanding of the events surrounding Columbus's discovery and settlement of the New World.

Complex Thinking Learning Targets: Historical Investigation

1. Your ability to identify and explain the confusion, uncertainty, or contradiction surrounding a past event.
2. Your ability to develop and defend a logical and plausible resolution to the confusion, uncertainty, or contradiction surrounding a past event.

Effective Communication Learning Targets

1. Your ability to communicate for a variety of purposes.
2. Your ability to communicate in a variety of ways.

Collaboration Learning Targets

1. Your ability to work with all of the students in your group to complete the project successfully.
2. Your ability to contribute good ideas and resources for presenting the findings to the class.
3. Your ability to do several different kinds of activities to help the group complete the project successfully.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (p. 60), by R. J. Marzano, D. Pickering, and J. McTighe, 1993, Alexandria, VA: Association for Supervision and Curriculum Development. Adapted by permission of McREL, 4601 DTC Blvd. #500, Denver, CO 80237.

FIGURE 12.5 **General rubrics for assessing collaboration and cooperation as students work in groups.**

Learning Target A: Works toward the achievement of group goals.

4 Actively helps to identify group goals and works hard to meet them.

3 Communicates commitment to the group's goals and effectively carries out assigned roles.

2 Communicates commitment to the group's goals but does not carry out assigned roles.

1 Does not work toward group goals or actively works against them.

Learning Target B: Demonstrates effective interpersonal skills.

4 Actively promotes effective group interaction and the expression of ideas and opinions in a way that is sensitive to the feelings and knowledge base of others.

3 Participates in group interaction with prompting. Expresses ideas and opinions in a way that is sensitive to the feelings and knowledge base of others.

2 Participates in group interaction without prompting or expresses ideas and opinions without considering the feelings and knowledge base of others.

1 Does not participate in group interaction, even with prompting, or expresses ideas and opinions in a way that is insensitive to the feelings and knowledge base of others.

Learning Target C: Contributes to group maintenance.

4 Actively helps the group to identify changes or modifications necessary in the group process and works toward carrying out those changes.

3 Helps identify changes or modifications necessary in the group process and works toward carrying out those changes.

2 When prompted, helps identify changes or modifications necessary in the group process, or is only minimally involved in carrying out those changes.

1 Does not attempt to identify changes or modifications necessary in the group process, even when prompted, or refuses to work toward carrying out those changes.

Learning Target D: Effectively performs a variety of roles within a group.

4 Effectively performs multiple roles within the group.

3 Effectively performs two roles within the group.

2 Makes an attempt to perform more than one role within the group but has little success with secondary roles.

1 Rejects opportunities or requests to perform more than one role in the group.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (pp. 87–88), by R. J. Marzano, D. Pickering, and J. McTighe, 1993, Alexandria, VA: Association for Supervision and Curriculum Development. Adapted by permission of McREL, 4601 DTC Blvd. #500, Denver, CO 80237.

after the group activities are completed, individuals prepare their own reports without assistance from the other group members. The combination approach is useful when a project is complex and requires the collaborative talents of several students to complete in a reasonable time frame, yet the learning target requires that individual students have the ability to prepare final reports, interpret results on their own, and so on. Research on cooperative learning indicates that students achieve most when the learning setting requires both group goals and individual accountability (Slavin, 1988). Assessment in this type of combined group and individual learning project requires assessing a group's joint success on the project as well as the degree to which individuals attained the learning targets.

Experiments

An **experiment** or *investigation* is an on-demand performance in which a student plans, conducts, and interprets the results of an empirical research study. The study focuses on answering specific research questions or on investigating specific research hypotheses. As defined here, experiments or investigations include a wide range of research activities that occur in both natural and social science disciplines. They include field and survey research investigations as well as laboratory and control-group experiments and may be conducted as individual or as group activities.

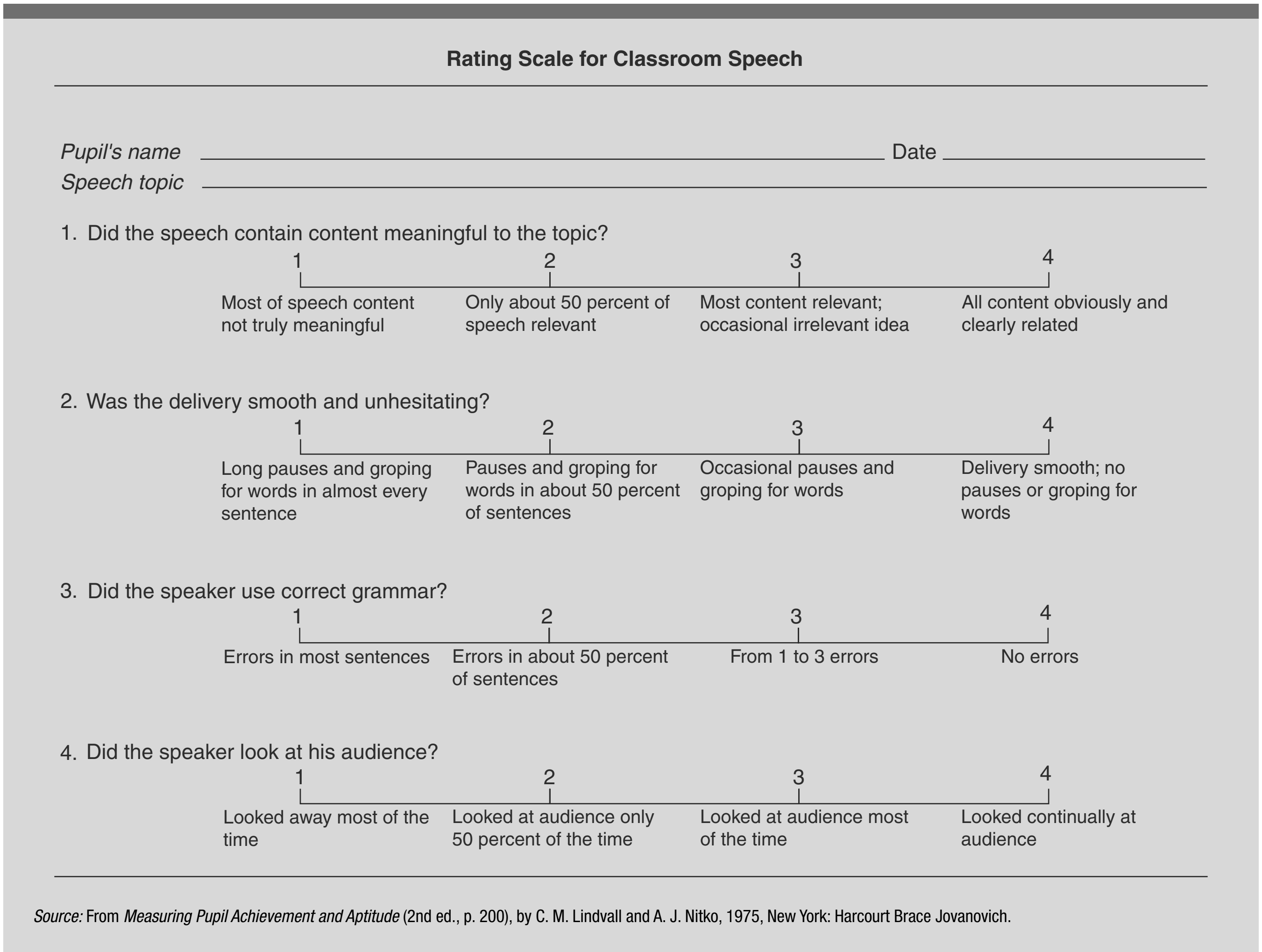
Experiments let you assess whether students use proper inquiry skills and methods. You can also assess whether students have developed proper conceptual frameworks and theoretical, discipline-based explanations of the phenomena they have investigated. To assess these latter aspects, focus on the quality of students' frames of reference, their mental representations of the problem they are studying, how well they plan or design the research, the quality of the questions or hypotheses they can specify, and the quality of explanations they offer for why the data relationships exist.

Oral Presentations and Dramatizations

Oral presentations permit students to verbalize their knowledge and use their oral skills in the form of interviews, speeches, or oral presentations. In language and language-arts curricula, many learning targets focus on style and communication skills rather than on the correctness of the content. Fluency of speaking a foreign language is an important learning target in some curricula. Another area in which oral presentations are especially useful is in speaking to a group. Figure 12.6 shows a simple scale for assessing the delivery of a classroom speech.

Debates are a special type of oral performance. A **debate** pits one student against another to argue issues logically in a formal exchange of views. Assessment focuses on the logical and persuasive quality of the argument and the rebuttals. Other *forensic activities* include poetry reading and oratories. **Dramatizations** combine verbalizations, oral and elocution skills, and movement performances. Students may express their understanding of fictional characters or historical persons, for example, by acting a role showing ideological positions and personal characteristics of these persons. For debates and other oral presentations that assess content knowledge as well as oral skills, it is usually a good idea to have separate rubrics for content and for oral presentation.

FIGURE 12.6 **Example of a simple rating scale for assessing the quality of a student's oral presentation.**



Rating Scale for Classroom Speech

Pupil's name ______________________________ Date ______________

Speech topic ____

1. Did the speech contain content meaningful to the topic?

1	2	3	4
Most of speech content not truly meaningful	Only about 50 percent of speech relevant	Most content relevant; occasional irrelevant idea	All content obviously and clearly related

2. Was the delivery smooth and unhesitating?

1	2	3	4
Long pauses and groping for words in almost every sentence	Pauses and groping for words in about 50 percent of sentences	Occasional pauses and groping for words	Delivery smooth; no pauses or groping for words

3. Did the speaker use correct grammar?

1	2	3	4
Errors in most sentences	Errors in about 50 percent of sentences	From 1 to 3 errors	No errors

4. Did the speaker look at his audience?

1	2	3	4
Looked away most of the time	Looked at audience only 50 percent of the time	Looked at audience most of the time	Looked continually at audience

Source: From *Measuring Pupil Achievement and Aptitude* (2nd ed., p. 200), by C. M. Lindvall and A. J. Nitko, 1975, New York: Harcourt Brace Jovanovich.

Simulations

Actors and "Standardized Patients" **Simulations** are on-demand events that happen under controlled conditions and attempt to mimic naturally occurring events. When the performance to be evaluated is the ability to interact with another person, an actor may be trained to play the role of the other person. Originally the **standardized patient format** was used to assess the clinical skills of medical candidates and practicing doctors. The actor is trained to display the symptoms of a particular disorder. Each medical candidate meets and interviews this standardized patient to diagnose the illness and to prescribe treatment. A panel of evaluators observes this interaction and assesses the candidate.

Computerized Adaptive Audiovisual Scenarios The combined technologies of video, CD-ROM, audio, and computers may be used to present realistic situations to students. A computer then evaluates students' responses to these presentations. If the situation presented is reasonably structured and the number of possible actions is limited, an **adaptive assessment task** can be built whereby a student's response to one situation will determine what the next presentation will be. For example, the media present a **scenario** to the student and ask a question or call for a decision. The student responds, and the presentation continues in a way that depends on the response. In this way each student receives a somewhat different scenario, depending on the individual choices of action.

Computerized Adaptive Text Scenarios This assessment format is similar to the adaptive audiovisual scenario format, except that text displays replace multimedia presentation.

Computerized Audiovisual Simulations With rapid advances in technology and software, multimedia simulations have become more realistic and complex. In middle school science, for example, computers can simulate hands-on investigations (Shavelson & Baxter, 1991). In a sow bug investigation, students investigate what the "computer sow bugs" do when simulated light and moisture are varied. Virtual reality technology offers more promise in this area. Flight simulators are examples of how technology combines with a sophisticated knowledge base of the conditions of the live performance and the consequences of different actions.

The advantages of this and the preceding formats are greater economy and consistency compared to real life, and the potential for computerized scoring (Jones, 1994). The further away from actual situations and real people the simulation gets, the less realistic and meaningful it is. Often, classroom teachers cannot use simulations because they are unavailable. These are the principal disadvantages of these formats.

ADVANTAGES AND CRITICISMS OF PERFORMANCE ASSESSMENTS

Advantages of Performance Assessment

Performance assessments have several advantages over other assessments (Hambleton & Murphy, 1992; Rudner & Boston, 1994; Shepard, 1991; Wiggins, 1990):

1. *Performance tasks clarify the meaning of complex learning targets.* Authentic performance tasks can closely match complex learning targets. When you present them to students and share them with parents, you make the learning goals clear through actual example.

2. *Performance tasks assess the ability "to do."* An important school outcome is the ability to use knowledge and skill to solve problems and lead a useful life, rather than simply to answer questions about doing.

3. *Performance assessment is consistent with modern learning theory.* Constructivist learning theory emphasizes that students should use their previous knowledge to build new knowledge structures, be actively involved in exploration and inquiry through tasklike activities, and construct meaning for themselves from educational experience. Most performance assessments engage students and actively involve them in complex tasks.

4. *Performance tasks require integration of knowledge, skills, and abilities.* Complex performance tasks, especially those that span longer periods, usually require students to use many different skills and abilities.

5. *Performance assessments may be linked more closely with teaching activities.* When your teaching requires students to be actively involved in inquiry and performance activities, performance assessments

are a meaningful component. Of course, this is not an advantage of performance assessment if your teaching is primarily teacher directed or uses lecture style.

6. *Performance tasks broaden the approach to student assessment.* Introducing performance assessment along with traditional objective formats broadens the types of learning targets you assess and offers students a variety of ways of expressing their learning. This increases the validity of your student evaluations.

7. *Performance tasks let teachers assess the processes students use as well as the products they produce.* Many performance tasks offer you the opportunity to watch the way a student goes about solving a problem or completing a task. Appropriate scoring rubrics help you collect information about the quality of the processes and strategies students use, as well as assess the quality of the finished product.

Disadvantages of Performance Assessments

Although performance assessments offer several advantages over traditional objective assessment procedures, they have some distinct disadvantages (Hambleton & Murphy, 1992; Miller & Seraphine, 1993; Rudner & Boston, 1994):

1. *High-quality performance assessments are difficult to create.* They need to match the outcome to be assessed but not add on other qualities (e.g., reading ability, dramatic skill, artistic flair) that are not part of the outcomes to be assessed. High-quality scoring rubrics are difficult to create, as well. They need to effectively capture all relevant characteristics, with descriptions that can be reliably used by students and teachers alike, and yet allow for the multiplicity of ways students might accomplish open-ended performance tasks.

2. *Completing performance tasks takes students a lot of time.* Even short on-demand paper-and-pencil tasks take 10 to 20 minutes per task to complete. Most authentic tasks take days or weeks to complete. If your assessments are not part of your instructional procedures themselves, this means either administering fewer tasks (thereby reducing the reliability of the results) or reducing the amount of instructional time.

3. *Scoring performance task responses takes a lot of time.* The more complex the performance and the product, the more time you can expect to spend on scoring. You can reduce the scoring time by crafting high-quality scoring rubrics. Holistic scoring is quicker than analytic scoring.

4. *Scores from performance tasks may have lower scorer reliability.* With complex tasks, multiple correct answers, and fast-paced performances, scoring depends on your own scoring competence. If two teachers use different frameworks, have different levels of competence, use a different scoring rubric, or use no scoring rubrics at all, they will mark the same student's performance or product quite differently. Inconsistent scoring is not only frustrating to a student, it also lowers the reliability and validity of the assessment results. Scorer reliability is especially problematic for portfolio assessment (Koretz, Stecher, Klein, & McCaffrey, 1994).

5. *Students' performance on one task provides little information about their performance on other tasks.* A serious problem with performance assessments is that a student's performance on a task very much depends on his or her prior knowledge, the particular wording and phrasing of the task, the context in which it is administered, and the specific subject-matter content embedded in the task (Lane et al., 1992; Linn, 1993; Shavelson & Baxter, 1991). This results in low reliability from the content-sampling point of view. You may have to use six or seven performance tasks to reliably evaluate a student for a learning target which implies that a student should be able to perform several quite different tasks under varied conditions and in several contexts.

6. *Performance tasks do not assess all learning targets well.* If a learning target focuses on memorizing and recalling, then objective format items such as short-answer, multiple-choice, matching, and true-false are better assessment choices. If your learning targets emphasize logical thinking, understanding concepts, or verbal reasoning, objective formats may still be a better choice than performance formats. They allow a much broader content to be assessed and can assess that broad coverage in less time. Further, objective formats are easier to score and the results from them are more reliable. A balanced assessment approach is recommended.

7. *Completing performance tasks may be discouraging to less able students.* Complex tasks that require students to sustain their interest and intensity over a long period may discourage less able students. They may have partial knowledge of the

learning target but may fail to complete the task because it does not allow them to use or express this partial knowledge effectively. Group projects may help by permitting peers to share the work, use each other's partial knowledge and differential skills, and motivate one another.

8. *Performance assessments may underrepresent the learning of some cultural groups.* Performance tasks will not wash away differences among cultural groups; in fact, they are likely to make such differences more apparent. Multiple assessment formats may improve this situation somewhat because they allow knowledge, skills, and ability to be expressed in different formats and media, thus allowing students with different backgrounds to express their achievement of the learning targets in different ways.

9. *Performance assessments may be corruptible.* As you use performance assessments, you will teach your students how to do well on them. This amounts to coaching them how to perform (often called "teaching to the test"). If your coaching amounts to teaching all aspects of your state's standards and your school's curriculum framework's learning targets, you are doing the right thing. However, if you focus primarily on only one aspect of the learning targets (e.g., how to write answers to constructed-response social studies items), you will lower the validity of your results.

In Chapters 2 and 11, we said that students' use of higher-order thinking is best assessed when they face new or novel tasks. Coaching tends to reduce the novelty of the task or change it from an "application" task to a "following-the-solution-strategy-the-teacher-taught-me" task. These types of coaching reduce the validity of your results because they do not assess the main intent of learning targets that want students to learn to solve new and ill-structured problems.

CREATING PERFORMANCE ASSESSMENTS

A systematic approach to creating performance tasks starts with being very clear about the performance you want to assess. Then, you design both a task and scoring scheme that match the intended performance. A well-designed performance task gives students the opportunity to apply their learning to a new situation. This shows them that learning must not be limited to repeating what teacher said. Design performance tasks to help students make connections between the skills and abilities they learned in separate subjects, and between "schoolhouse" learning and real-world activities. Share your scoring rubrics with students to clarify the learning targets for them.The more students understand the skills and abilities they should use, the better able they are to identify where they should focus their practice and study efforts.

STAGE ONE: BE CLEAR ABOUT THE PERFORMANCE TO ASSESS

You may decide that two or three learning targets can be assessed by the same complex performance assessment. Some learning targets may cut across curricula (e.g., effective communication). Select only those learning targets that can and should be assessed by performance tasks. Make sure the performance assessment you design fits into your assessment plan, along with tests and other assessments for other learning targets. The goal is a balanced assessment of worthwhile learning targets of both knowledge and skills.

You should answer the following questions to identify a conceptual framework and the important achievement dimensions or criteria to assess (Herman, Aschbacher, & Winters, 1992):

- What are the characteristics of high-quality achievement (e.g., good writing, good problem solving, good collaboration, good scientific thinking, etc.)? What evidence should I look for to decide if a student has produced an excellent response?
- What are the important characteristics of the learning target that I should assess?
- What is it about the students' responses that distinguish the poor, acceptable, and excellent work?
- Are there samples of student work (excellent and poor) that I can contrast to identify the characteristics that differentiate them?
- Does my school district, state assessment program, a national curriculum panel, or a professional society have examples of rubrics or curriculum frameworks that show standards and criteria?
- Are there any suggestions in teachers' magazines, state teacher's newsletters, professional journals, or textbooks?

Define Quality Levels

Each achievement dimension you assess actually represents a continuum of educational growth. Different students will attain different levels of achievement on each dimension. Further, one student may perform with high competence on some dimensions but with less competence on others. Thus, part of creating your performance task is to define an achievement scale for each dimension. You define this scale by spelling out the different degrees of achievement—from low to high—on each dimension. This continuum forms the basis for crafting scoring rubrics, which we discuss later in this chapter.

Evaluate the Achievement Dimensions You Select

You may wonder whether the achievement dimensions you selected are appropriately stated. The checklist that provides criteria for evaluating these achievement dimensions.

✔ **Checklist**

A Checklist for Judging the Quality of the Achievement Dimensions You Intend to Use to Evaluate Students on a Performance Task

Ask these questions of every achievement dimension you intend to assess. If you answer no to one or more questions, revise the description of the achievement accordingly.

1. *Do the achievements you are assessing have significance within the broader context of the curriculum and real-world applications?* Your achievement dimensions for your task should specify only the most important components; they should include high-level content learning targets and reflect several lifelong or real-world learning targets, including complex reasoning, information processing, effective communication, and, where appropriate, habits of mind and cooperative learning targets.
2. *Do the achievements you are assessing have authenticity and fidelity to the way the task should be performed outside the assessment context?* Your achievement dimensions should be stated in a way that they would apply equally if the student were to perform a similar task in the real world. They should reflect factors such as the resources typically available when the task is performed in the real world, as well as specifying the types of structure and assistance (i.e., scaffolding) a student would have available to complete the task in the real world.
3. *Have you applied a general achievement framework to your scoring rubrics for this task?* Your rubrics, although specific to the performance task you are using, should fit into a general scheme or general rubric framework so it is easy to see how a student has performed over several similar tasks belonging to the same category, but under different conditions or at different times. This type of general framework will make it easier for you and other teachers to apply the rubric consistently across different tasks within the same curriculum.
4. *Do your achievement dimensions fit within a broader framework of educational competence?* Your achievement dimensions should be (a) stated in a sound educational development way so it is easy to see as extending from novice to expert performance, (b) located within a broader framework but in a way that is appropriate for the grade and age levels of the students you are assessing, and (c) worded to describe the performance expected at each level (as contrasted with being stated as values such as "poor").
5. *Are your achievement dimensions easy for students, parents, and teachers to understand?* Your achievement dimensions should be stated in clear language so students, parents, teachers, and the community easily understand them. You may want to have a plain language version and a technical language version, the former for students and parents and the latter for other teachers in your field.
6. *Are your achievement dimensions useful for pointing to the ways students can improve?* Your achievement dimensions should focus on those features of performance that students can improve. Your dimensions should communicate to students (and others) what they need to concentrate on to improve.

Source: Based on criteria and ideas in Quellmalz (1991).

Should You Assess Process, Products, or Both?

Sometimes a learning target asks students to demonstrate a process or a procedure. Focus your assessment on the *process* students use if you taught students to use a particular procedure for which you can specify steps and accurately assess the extent to which your students follow the accepted procedure(s). Focus your assessment on the process students use if most of the evidence about the students' achievement of the learning target is found in the way the performance is carried out, and little or none of the evidence you need to evaluate the students is present in the product itself. Here are some examples:

Examples

Examples of Learning Targets that Require Students to Demonstrate a Certain Process

1. Use the long-division algorithm.
2. Use the posted safety procedures when handling laboratory chemicals.

In some cases, the learning target permits several correct processes. In a mathematics curriculum, for example, a learning target may ask students to learn several different procedures for division rather than a single correct algorithm.

At other times, even though you teach a specific process, the curriculum framework clearly implies that the major focus is the product students produce. Focus your assessment on the *product* students produce if most or all of the evidence about their achievement of the learning targets is found in the product itself, and little or none of the evidence you need to evaluate students is found in the procedures they use or the ways in which they perform. Focus your assessment on the product students produce if there are multiple good ways to produce a high-quality product, and the exact method or sequence of steps does not make much difference as long as the product is good. Here are two examples:

Examples

Examples of Learning Targets that Require Students to Produce a Product

1. Write haiku poems based on everyday experience.
2. Prepare a research term paper on the causes of volcanic eruptions.

There may be several equally good methods for completing such tasks, but the focus is on the result or products.

Sometimes *both product and process* are of equal importance. For example:

Example

Examples of Learning Targets that Require Students to Produce a Product by Performing a Certain Process

1. Writing a research term paper by following certain steps you outline.
2. Using the long-division algorithm, solve 90% of the problems presented in the chapter quiz.

As you create an assessment task, be sure that performing the task is within the students' ability range. Performance assessments differ depending on the students' educational level and the mix of general scholastic ability in your class. Further, some students with disabilities may need to have the tasks modified before they can participate in the performance assessment (see Chapter 5). In addition, tasks should suit class size. The more students you have in your class, the less elaborate the performance assessment tasks you can set. The fewer the number in your class, the more time you have per student for scoring. Your assessment planning should reflect the realities of your classroom.

STAGE TWO: DESIGNING PERFORMANCE TASKS

When you have a clear understanding of the achievement you want to assess, the next step is to design the task(s) that will assess this achievement. The types of tasks you craft will depend on the learning targets you are assessing. Some targets imply that the tasks should be structured; others require unstructured tasks; tasks can be structured in various ways. The questions you must answer as you design your tasks include the following:

- What ranges of tasks do the learning targets imply?
- Which parts of the tasks should be structured, and to what degree?
- Does each task require students to perform all the important elements implied by the learning targets?
- Do the tasks allow me to assess the achievement dimensions I need to assess?
- What must I tell students about the task and its scoring to communicate to them what they need to perform?
- Will students with different ethnic and social backgrounds interpret my task appropriately?

Create Meaningful Tasks

The tasks you craft should be meaningful to the students. This lets students become personally involved in solving a problem or doing well on the task. If possible, choose a situation or task that is likely to have personal meaning for most of your students. Carefully blend the familiar and the novel so students will be challenged but not frustrated by the task. Choose situations or tasks that assess whether students have the ability to transfer their knowledge and skills from classroom activities and examples to similar but new (for them) formats (Baron, 1991).

Create a draft of your task with your learning targets clearly in mind. Consider whether you want a group or individual task. Consider the content knowledge and skills required, and the kind of thinking (analysis or synthesis, for example) required to master the learning target.

Trying out assessment tasks (whether performance or traditional paper-and-pencil) before using them is next to impossible for classroom teachers. You can, however, have your colleagues review and criticize your tasks before you use them. The next best thing to actual student tryouts can and should be done: After you use an assessment task, use the information you obtain about flaws in the task or in the rubrics to revise the task or rubric; then reuse the task and rubric next year with a new class of students.

Things to Control to Craft Valid Tasks

Tasks assessing the same content learning target can differ from one another. These differences make some tasks useful for assessing different types of lifelong learning targets. Figure 12.7 shows five properties of a task that you must control to produce a well-designed task. Study your task's learning target to decide how to control these properties to make your task more valid.

Time Needed to Complete the Work Some learning targets can be assessed in a relatively short period of 15 to 40 minutes. For example, the ability to work in groups, write an essay or an explanation,

FIGURE 12.7 **Properties of tasks that you can vary to better align students' performance with the requirements of the achievement dimensions and learning targets.**

Task property	Variations in the task requirements
Time to complete the task	*Short tasks* can be done in one class period or less. *Long tasks* require a month or more, and work may need to be done outside class.
Task structure provided	Structure may vary in: *Problem definition:* High structure means you carefully define the problem the students must solve. Low structure means students are free to select and define the problem. *Scaffolding:* High structure means students are given lots of guidance or directions in how to begin a solution and what materials to use. Low structure means students have little or no guidance and must decide for themselves. *Alternate strategies:* High structure means there are very few correct or appropriate pathways to get to the correct answer. Low structure means there are many correct or appropriate approaches to get an acceptable answer. *Alternate solutions:* High structure means there is a correct answer to the task. Low structure means there is no single correct answer to this task.
Participation of groups	The task may require: *Individual work* only throughout all phases of performance. *Mixed individual and group work* in which some of the performance occurs in groups and some is strictly individual effort. *Group work* only throughout all phases of performance.
Product and process focus	The task may require: *Process assessment* only in which the students' performance of the steps and procedures and not the outcome are observed and evaluated. *Both process and product assessment* in which both the steps and the concrete outcome (product) are evaluated. *Product assessment* only in which only the concrete product or outcome is evaluated.
Performance modality	The task may require: *A single modality* in which the performance is limited to one mode (e.g., oral, written, wood model, etc.). *Multiple modality* in which the performance must be done in several modes (e.g., do both a written and an oral report).

Source: Based on ideas in Davey & Rindone (1990).

plot a graph, or carry out simple experiments can be assessed with short tasks. Many learning targets and dimensions, however, necessitate that students complete long tasks. For example, doing an opinion survey and writing it up, building a model town, and developing complex plans for community action require a month or more, and much of the work may need to be done outside class. Task time limits must match the intent of the learning target and dimensions rather than your own convenience—if your goal is to use the results to make valid interpretations about how well a student has achieved that learning target.

Task Structure It may be misleading to talk about structured versus unstructured performance because you can **structure a task** in various ways (Davey & Rindone, 1990), including the way you define the problem, scaffold the instructions, require alternate strategies, and require alternate solutions. At one extreme your task may *define a problem* for students to solve (structured); at the other extreme you may require the students themselves to identify what the problem is (unstructured or ill-defined).

Scaffolding is the degree of support, guidance, and direction you provide the students when they set out to complete the task. You may suggest how to attack the problem, what books or material to use, and the general nature of the end product you require. These directions and guidance statements add structure to the task. Less scaffolding means less structure.

If your task can be performed or solved using only one or two procedures or strategies, it has fewer **alternate solution strategies** and is more structured in this respect. Unstructured alternatives mean that there are a great many equally correct pathways to the correct answer or to producing the correct product. A similar analysis applies to the solution or the product itself: A task is unstructured in this respect when it has many correct or *acceptable solutions or products*. Just how a task may vary in these elements is shown in the following example:

Example

Example Showing How Controlling Properties can Change a Performance Task

Assume that students have been asked simply to build a scale model of the solar system. As far as problem definition goes, this task has fairly high structure—you have a specified goal to meet. However, there is very little scaffolding—students are not told what materials to use, or what proportions to use, or where to get information on the planetary distances and orbits. There are a lot of alternative pathways to the solution—consider the fact that no two models will look exactly alike, and will vary in terms of materials used, scale employed, special features included (such as neurons, orbital speeds, etc.), and in a way there's one best solution, a perfectly scaled model of the solar system. (Davey & Rindone, 1990, p. 5)

Participation of Groups Learning targets and the dimensions guide your task construction. If your learning targets call for cooperative or collaborative learning (or using other group-based skills), you should set a task using, at least in part, group activities.

Product and Process If you want to assess process, you need to do the assessing while the students are performing. You may take away a product, on the other hand, and evaluate it at your convenience. Further, you cannot assess cognitive processes (mental activities) directly, only indirectly through some intermediate or "partial" products. For example, you can ask students to tell you or to write what they were thinking about while they were doing the task. Or you may ask them to record the early drafts they made and ideas they used.

Your indirect assessments of students' mental activities and thinking processes depend on the students having abilities other than those required to complete the task. They depend, for example, on the accuracy of the students' memories, their skills in understanding the thinking processes they used, and their abilities to describe these thinking processes orally or in writing. Because you assess cognitive processes only indirectly, your inferences and judgments about how well students use them—that is, the validity of your evaluation of students' use of cognitive processes—might be weak. Other processes, such as group processes and behavior that occurs in a sequence of steps, are more directly assessed because you can observe them directly.

Response Mode Some learning targets specify that students should be able to communicate their knowledge in several ways, solve a problem using several methods, or express themselves in a variety of modalities. For example, students may be required to use written or oral reports, posters, presentation software, or brochures. You should not use multiple response modes on a whim, however. Align the modalities with the learning targets, state standards, and curriculum framework. Also, use alternate modes to accommodate students with disabilities or cultural differences if the mainstream, single mode is not appropriate for them.

Make the Task Clear to Students

When crafting performance tasks, you write the learning target(s), the criteria by which you will evaluate performance, and the instructions for completing the task. Task wording and directions should depend on the educational maturity of your students. Clearly state the time limits and the conditions under which you want the task done. Be sure students understand how long a response you are expecting. Share with students the rubrics you will use to assess their performance.

When students misinterpret the task, you cannot validly interpret their assessment results in the same way as you do those students who interpreted the task correctly. Students from certain ethnic, linguistic, or gender groups may not interpret your wording as you expect them to (Duran, 1989; Lane et al., 1992).

Number of Tasks

As a general rule, the fewer the number of tasks, the fewer learning targets you can assess, the lower the score reliability, and the lower the validity of your interpretations. The number of performance tasks to include in your assessment depends on several factors; however, some of these factors you cannot control. You need to resolve the following issues to decide on the appropriate number of tasks:

1. *Crucial decisions.* Certification, promotion, and graduation are examples of very crucial decisions. Assessments for crucial decisions such as these, in which the consequences of failing are severe, are high-stakes assessments. These assessments require more tasks and longer assessment times to gather sufficiently reliable information. High-stakes decisions should not depend on information from only one assessment session. Letter-grade assignments may be less crucial a decision than certification, but grades for a term or a marking period should not be based on a single assessment, either. Daily instructional decisions for formative evaluation of learning can be easily changed if wrong decisions are made. These less crucial decisions can be based on lower-quality information if need be.

2. *Scope of your assessment.* How much instruction are you covering with this assessment—a unit or only one lesson? How much content is covered in a unit? The broader the scope of your assessment, the more tasks you will need.

3. *Mixture of assessment formats.* If you mix objective formats with performance tasks, you will be able to cover more aspects of the learning targets, balance your assessment, and broaden your assessment scope. In this case, you may need fewer performance tasks because your assessment scope will be broader than if you used performance tasks alone.

4. *Complexity of the learning target.* A complex learning target requires integration of many skills and abilities and may need to be performed over a long time. In this case, practicality limits the number of tasks of this type you may give. However, because more time is devoted to one (or at most a few) such tasks, the information may be quite reliable. Nevertheless, the scope or span of your assessment may not be very broad. This could present a validity problem.

5. *Time needed to complete each task.* As a practical matter, you can administer only a few tasks during a typical class period. Estimate how much time one task will take students to complete, and divide this into the length of the class period to determine the maximum number of tasks possible.

6. *Time available for the total assessment.* You may be willing to devote more or less than one period to assessment. The number of tasks may shrink or expand depending on the available time.

7. *Diagnostic detail needed.* If you need a lot of detail to diagnose a student's learning or conceptual problems, you need to craft tasks that provide this rich detail. This usually means fewer tasks, more detailed performance, and more detailed scoring of the responses. If you assess many students for diagnostic purposes, practicalities of time for performance and scoring will usually limit you to only a few tasks per student.

8. *Available human resources.* If you have an aide or a parent to help you administer or score the

assessments, this may free up some time so that you can give a few more tasks. You can also teach students to score the assessments; although this is educationally useful, it is unlikely to lead to increasing the number of tasks.

Evaluate Your Performance Tasks

The more important suggestions for improving performance tasks are shown in the checklist that follows. You can use this checklist to evaluate individual performance tasks.

✔ **Checklist**

A Checklist for Judging the Quality of Performance Tasks

Ask these questions of every performance task you design. If you answer no to one or more questions, revise the task accordingly before administering it to students.

1. Does the task focus on an important aspect of the unit's learning targets?
2. Does the task match your assessment plan in terms of performance, emphasis, and number of points (marks)?
3. Does the task actually require a student to *do* something (i.e., a performance) rather than requiring only writing about how to do it, or simply to recall or copy information?
4. Do you allow enough time so all of your students can complete the task under your specified conditions?
5. *If this is an open-response task,* do your wording and directions make it clear to students that they may use a variety of approaches and strategies, that you will accept more than one answer as correct, and that they need to fully elaborate their response?
6. *If the task is intended to be authentic or realistic,* do you present a situation that your level of students will recognize as coming from the real world?
7. *If this task requires using resources and locating information outside the classroom,* will all of your students have fair and equal access to the expected resources?
8. Do your directions and other wording:
   a. define a task that is appropriate to the educational maturity of your students?
   b. lead all students, including those from diverse cultural and ethnic backgrounds, to interpret the task requirements in the way you intend?
   c. make clear the purpose or goal of the task?
   d. make clear the length or the degree of elaboration of the response that you expect?
   e. make clear the bases on which you will evaluate the responses to the task?
9. Are the drawings, graphs, diagrams, charts, manipulatives, and other task materials clearly drawn, properly constructed, appropriate to the intended performance, and in good working order?
10. If you have students with disabilities in your class, have you modified or adapted the task to accommodate their needs?

STAGE THREE: DESIGNING SCORING SCHEMES

Several useful ways to record your assessments of students' performance are briefly described in Figure 12.8. Although each of the ways listed in the figure has a special use, rubrics, checklists, and rating scales are most frequently used with performance tasks. Suggestions for developing rubrics, checklists, and rating scales are detailed in the following sections.

Rubrics

Rubrics not only improve scoring consistency, they also improve validity by clarifying the standards of achievement you will use to evaluate your students. As you craft scoring rubrics and ways of recording results, you will need to address questions such as:

- What important criteria and learning targets do I need to assess?
- What are the levels of development (achievement) for each of these criteria and learning targets?
- Should I use a holistic or an analytic scoring rubric?
- Do I need to use a rating scale or a checklist as my scoring scheme?
- Should my students be involved in rating their own performance?
- How can I make my scoring efficient and less time-consuming?
- What do I need to record as the result of my assessments?

FIGURE 12.8 **Some useful methods of recording students' responses to performance tasks.**

Recording method	Description	Recommended use	Example of uses
Anecdotal records	You observe the performance and write a description of what the student did.	These are primarily useful for keeping records of unanticipated or naturally occurring performances. Usually you can record only one student at a time.	A student shows unusual insights into current events and you want to keep a record of these to put into his or her portfolio or to recommend the student for a summer program for leadership.
Behavior tallies	You create a list of specific behaviors of which you want to keep a record for a student. As you observe the performance you tally how many times each behavior occurs. The list is usually limited to only a few behaviors.	These are primarily useful for well-defined lists of behaviors that you can expect to occur frequently. They also may be useful to keep track of undesirable behaviors.	As a communications teacher, you keep track of how often a student uses "uh-h-h" when speaking in front of the class.
Checklists	You create a list of specific steps in a procedure or specific behaviors. You check each behavior that occurs. The list may be long.	These are primarily useful if the behaviors are in a sequence or if all the subtasks that make up the complete performance can be listed.	You are a science teacher and want to be sure that each student performs the steps in setting up a microscope properly. You are an automotive shop teacher and want to be sure that each student properly completes all the tasks necessary to change the oil in a car.
Rating scales	You create standards or criteria for evaluating a performance. Each standard has levels of competence, and you rate students according to how well they performed each standard as they complete the task.	These are especially useful if each standard can be judged according to the level or the degree of quality rather than as simply being present or absent.	You are an art teacher and rate each student's painting on its composition, texture, theme, and technique. You are a mathematics teacher and rate a student's problem solution according to how well the student demonstrates mathematical knowledge, uses a good strategy to solve the problem, and communicates his or her explanation of the solution in writing.
Rubrics	In some ways, rubrics are a type of rating scale. By convention, "rubrics" usually refers to a scale on which each level has a complete description of performance quality, and "rating scale" usually refers to a scale on which the levels are anchored with level or degree descriptions such as "frequently, occasionally, never" or "to a great degree, somewhat, not at all," and the like	These are useful for classroom instruction and assessment. Rubrics place descriptions in the hands of students, who can use them to produce work and to monitor their own work. Teachers can use them to clarify learning targets at the beginning of a lesson and to evaluate achievement at the end of a lesson.	You are a social studies teacher and use rubrics to evaluate a term paper according to quality of thesis, accuracy and completeness of content supporting the thesis, and quality of written presentation.

Rubrics can be categorized according to whether they use one scale or several and according to whether the descriptions of work quality are general (i.e., can be applied to many different tasks) or specific to the task, essay, or assignment. An **analytic scoring rubric** (also called *scoring key, point scale,* or *trait scale*) requires you to evaluate specific dimensions, traits, or elements of a student's response. A **holistic scoring rubric** (also called global, sorting, or rating) requires you to make a judgment about the overall quality of each student's response. **General rubrics** (also called *generic rubrics*) describe performance quality in general terms so the scoring can be applied to many different tasks. **Task-specific rubrics** describe performance quality in terms that reference the specific assignment. Note that whether a rubric is analytic or holistic is independent of whether it is general or task-specific. Rubrics can be described on both factors. For example, the writing rubrics in Appendix H are general (generic) and analytic.

Analytic Scoring Rubrics An analytic scoring rubric requires that you list the major criteria of good work (sometimes called *dimensions* or *traits*) and prepare a rubric for each of these criteria. You decide the number of points to award to students for each criterion. An example of an analytic, task-specific scoring rubric for a restricted-response essay was presented in Chapter 10. Examples of generic analytic rubrics for evaluating collaboration and cooperation were presented in Figure 12.5. The scales may all be of equal weight, or you may decide one or more of the aspects of performance is worth more points.

Usually students' responses will match the scoring rubric to various degrees. Assigning a rubric level to particular student work is like a "choose the best answer" type of multiple-choice question. The score is the one whose description most closely matches a student's work. The top and bottom of a rubric scale are usually easier categories to decide than the middle. When you match student work to rubric levels in an inconsistent way you lower the reliability of the scoring process.

Holistic Scoring Rubrics Holistic scoring is appropriate for extended response subject-matter essays or papers involving a student's abilities to synthesize and create when no single description of good work can be prespecified. It is also appropriate for final exams or projects where giving feedback to students is not a consideration. States that do large-scale assessment of either writing or subject-matter essay responses often prefer holistic scoring. The large numbers of papers to be marked often precludes the detailed scoring required by analytic rubrics. An example of a holistic, task-specific scoring rubric for a restricted-response essay was presented in Chapter 10.

To create holistic rubrics, you still need to identify the criteria for good work on which your scoring will be based. The difference is that for analytic rubrics, descriptions of levels of performance on each criterion are considered separately. For holistic rubrics, levels of performance on all criteria are considered simultaneously. The description that best fits the student work identifies the score to be given.

Decide beforehand on the number of categories of the overall quality of the work into which you will sort the students' responses to each question. Usually, you can use between three and five categories, such as A, B, C, D, and F; distinguished, proficient, apprentice, and novice; or 4, 3, 2, and 1. A particularly important point in deciding the number of categories is to be sure they correspond to your school's grading system. If your school uses grades A through F, for example, then you need five categories. Using only three quality levels in a scoring rubric will make your student evaluations unnecessarily complicated.

After deciding on the number of categories, define the *quality* of the papers that belong in each category. This means, for example, describing what constitutes an A performance, a B performance, and so on. It is best to revise your scoring rubric after you have tried out the draft version on several performances (papers, assignments, projects). Trying out the rubric will allow you to identify parts of it that are problematic, then add qualities of student responses that you may have failed to include in the original draft. After categorizing all of the students' work, you should reexamine the performances within categories to be sure they are enough alike in quality to receive the same grade or quality rating.

A refinement that will help you use the rubrics more reliably, and make them even easier to use the next time, is to select *specimens* or **exemplars** that are good examples of each scoring category. You can then compare the current students' answers to the exemplars that define each quality level. Finally, you decide into which category to place them.

An alternative way of implementing holistic scoring is to consider all the papers, projects, or assignments and compare one with another to decide which are the best, the next best, and so on. This will result in a rough ranking of all the papers. The best-ranked papers are placed in the highest category, the next best in the second category, and so on. This approach, however, does not work very well with a large number of students, and it is inconsistent with a criterion-referenced approach to teaching, which bases instruction on learning objectives and assessment on the degree to which each student met the objectives.

Annotated Holistic Scoring Rubrics Some educators have successfully used a third type of scoring rubric that is a hybrid of the analytic and holistic rubrics. The **annotated holistic rubric** is an approach that uses holistic scoring but adds feedback to students on a few of the traits in a way similar to the analytic scoring. With this approach, quality levels are defined and the papers are scored holistically. After reaching a holistic judgment, you write on the student's paper very brief comments, based on the prespecified traits, that point out one or two strengths of the response and one or two weaknesses. You write only about what led you to reach your holistic judgment of the paper.

General (Generic) Rubrics General (generic) rubrics use descriptions of work that apply to a whole family or set of assignments. General rubrics for writing, math problem solving, science laboratory work, analyzing literature, and so on are important instructional as well as assessment tools. As students practice and perform many different learning targets in a subject throughout the school year, their learning improves if they apply the same general evaluation framework to all of the same type of work in that subject. Some research evidence supports the idea that when students routinely use general, but analytic, rubrics in the classroom, their achievement improves (Khattri, Reeve, & Adamson, 1997). The Oregon writing assessment rubric in Appendix H is an example of this type of generic, analytic rubric.

A general (generic) scoring rubric contains guidelines for scoring that apply across many different tasks of a similar type (for example, writing, or math problem solving), not just to one specific instance of that kind of task. The general rubric can serve as an overall framework for developing more specific rubrics, or it can be used as is. Following is an example of a general scoring guide for assessing the content learning target of any high school history task similar to the one we presented previously. A task centered on the Gulf War and President Bush, for example, would be an alternative task that could assess this content learning target.

Example

General Rubric That Could Be Applied to a High School History Task

Content learning target being assessed:
Understands that war forces sensitive issues to the surface and causes people to confront inherent conflicts of values and beliefs.

4. Demonstrates a thorough understanding of the generalized concepts and facts specific to the task or situation. Provides new insights into some aspect of that information.
3. Displays a complete and accurate understanding of the generalizations, concepts, and facts specific to the task or situation.
2. Displays an incomplete understanding of the generalizations, concepts, and facts specific to the task or situation.
1. Demonstrates severe misconceptions about the generalizations, concepts, and facts specific to the task or situation.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (pp. 29–30), by R. J. Marzano, D. Pickering, and J. McTighe, 1993, Alexandria, VA: Association for Supervision and Curriculum Development. Adapted by permission of McREL, 4601 DTC Blvd. #500, Denver, CO 80237.

Task-Specific Rubrics A task-specific scoring rubric is a scoring scale that applies the general scoring framework to a particular task. Carefully applying the general scoring framework to craft a specific scoring rubric ensures that your specific rubric assesses students in a way that is aligned with the general scoring framework. This process would be very helpful when you use the state's general rubric to develop a specific rubric for your classroom because it helps you align your class assessments with the state standards. The example that follows is a specific scoring rubric that a history teacher used to assess students' performance on the high school history decision-making task presented earlier in Figure 12.4. It uses the preceding general scoring rubric as a framework.

Example

Specific Rubric That Could Be Applied to the High School History Task

Content learning target being assessed:
Understands that war forces sensitive issues to the surface and causes people to confront inherent conflicts of values and beliefs.

4. Demonstrates a thorough understanding of the generalization that war forces sensitive issues to the surface and causes people to confront inherent conflicts of values and beliefs. Provides new insights into people's behavior during wartime.
3. Displays a complete and accurate understanding of the generalization that war forces sensitive issues to the surface and causes people to confront inherent conflicts of values and beliefs.
2. Displays an incomplete understanding of the generalization that war forces sensitive issues to the surface and causes people to confront inherent conflicts of values and beliefs. Has some notable misconceptions about this generalization.
1. Demonstrates severe misconceptions about the generalization that war forces sensitive issues to the surface and causes people to confront inherent conflicts of values and beliefs.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (pp. 29–30), by R. J. Marzano, D. Pickering, and J. McTighe, 1993, Alexandria, VA: Association for Supervision and Curriculum Development. Adapted by permission of McREL, 4601 DTC Blvd. #500, Denver, CO 80237.

You can use a general framework to develop scoring rubrics specific to a particular task. In this way you are applying the same general framework in a consistent manner to each new performance task. The reliability and validity of your scores improve when you use a general scoring framework as a guideline for specific scoring rubrics.

Advantages and Disadvantages of Different Types of Rubrics Different scoring approaches are not interchangeable. They serve different purposes for scoring your students' performance. Figure 12.9 gives advantages and disadvantages for each type of rubrics.

A clear advantage of the analytic scoring rubric is that it provides you and your students with much more detail about their strengths and weaknesses. If you use an analytic scoring rubric, take advantage of this added information to enhance your teaching and to give students guidance concerning what they need to do to improve. For example, you could identify which elements or parts of the entire class's answers are weakest and direct your reteaching to that aspect. You will also be able to give your students specific feedback about those parts of the answer on which they did well.

Holistic scoring rubrics are easier to use and take less time per student. They permit an overall evaluation, which allows the rater to report a general impression over all aspects of the performance. An analytic scoring rubric is more time-consuming to use because the rater must look for and separately rate each component of a performance. This level of detail is useful when your focus is diagnosis or helping students understand your expectations for each part of the performance. This may be especially useful for helping students learn, even if it is more time-consuming. Using a general, analytic trait rubric (e.g., the type illustrated in Appendix H) in a consistent way throughout the entire year may improve learning if students understand it and if they receive feedback linked to it.

The annotated holistic scoring rubric is a restricted combination of the holistic and analytic rubrics. The additional analytic feedback is restricted to only a few characteristics, which do not change the initial holistic rating. The advantage is that it allows you to rate the papers quickly and to support your rating with a few salient points. These points give feedback to students but may not be useful for diagnosis. In order to provide a complete diagnosis and feedback, you still need analytic rubrics that rate each component of the performance separately.

Note that holistic and analytic scoring rubrics probably assess a student's performance differently (Taylor, 1998). Analytic trait scoring may be more valid if it allows you to evaluate several dimensions of performance as well as how the student integrates those dimensions when performing the task.

Task-specific scoring rubrics cannot be shared with students ahead of time. They contain specific information about the responses the students are expected to make, for example, "answers" to problems students are to solve, or lists of facts or concepts students should provide. And you obviously have to come up with a new rubric for each task. However, task-specific rubrics are very useful for some purposes. They make for reliable and efficient scoring of essay questions or show-the-work problems on exams. This is probably their best use. Because of the instructional and formative assessment advantages, general (generic), analytic rubrics are the kind you aim to use whenever students are involved in the assessment process—which should be most of the time.

Designing Scoring Rubrics: Before You Begin Scoring rubrics are necessary for all of the performance assessment methods described in this and the previous chapter, including projects and portfolios, and for scoring essays and show-the-work

FIGURE 12.9 **Advantages and disadvantages of different types of rubrics.**

Type of rubric	Definition	Advantages	Disadvantages
Holistic or Analytic: One or Several Judgments?			
Analytic	■ Each criterion (dimension, trait) is evaluated separately.	■ Gives diagnostic information to teacher. ■ Gives formative feedback to students. ■ Easier to link to instruction than holistic rubrics. ■ Good for formative assessment; adaptable for summative assessment; if you need an overall score for grading, you can combine the scores.	■ Takes more time to score than holistic rubrics. ■ Takes more time to achieve inter-rater reliability than with holistic rubrics.
Holistic	■ All criteria (dimensions, traits) are evaluated simultaneously.	■ Scoring is faster than with analytic rubrics. ■ Requires less time to achieve inter-rater reliability. ■ Good for summative assessment.	■ Single overall score does not communicate information about what to do to improve. ■ Not good for formative assessment.
Description of Performance: General or Task-Specific?			
General	■ Description of work gives characteristics that apply to a whole family of tasks (e.g., writing, problem solving).	■ Can share with students, explicitly linking assessment and instruction. ■ Reuse same rubrics with several tasks or assignments. ■ Supports learning by helping students see "good work" as bigger than one task. ■ Support student self-evaluation. ■ Students can help construct generic rubrics.	■ Lower reliability at first than with task-specific rubrics. ■ Requires practice to apply well.
Task-specific	■ Description of work refers to the specific content of a particular task (e.g., gives an answer, specifies a conclusion).	■ Teachers sometimes say using these makes scoring "easier." ■ Requires less time to achieve inter-rater reliability.	■ Cannot share with students (would give away answers). ■ Need to write new rubrics for each task. ■ For open-ended tasks, good answers not listed in rubrics may be evaluated poorly.

problems. In Stage One of the designing process, you identified achievement dimensions and a scale of progress for each dimension, from very low progress to very high progress. To create a scoring rubric you need to refine these descriptions of performance levels to be sure they are clear. You may associate each level with a numerical value. Alternately, you may associate each level with a qualitative description such as novice, apprentice, proficient, and distinguished. Describe the characteristics of a student's performance that distinguish one achievement level from another, because these descriptions anchor the scale at each level.

Crafting Scoring Rubrics: The Top-Down Approach The top-down approach begins with a conceptual framework that you can use to evaluate students' performance to develop scoring rubrics. Follow these steps:

Step 1. Adapt or create a conceptual framework of achievement dimensions that describes the content and performance that you should assess. Develop a detailed outline that arranges the content and performance to identify what you should include at each level of each dimension in the general rubric.

Step 2. Write a general scoring rubric that conforms to this detailed outline and focuses on the important aspects of content and process to be assessed across different tasks. The general rubric can be shared with students. It can be used as is to score student work, or it can be used to craft specific rubrics.

Step 3. Craft a specific scoring rubric for the specific performance task you are going to use.

Step 4. Use the specific scoring rubric to assess the performances of several students; use this experience to revise the rubric as necessary.

In the top-down approach you need a framework-based organization to develop a rubric. Thus, Steps 1 and 2 may be difficult to achieve on your own and may require you to work with groups of teachers.

Figure 12.10 shows an example of a holistic scoring rubric for middle school mathematics that has been organized around a three-part conceptual framework: mathematical knowledge, strategic knowledge, and communication (Lane, 1992). This three-part organization helps define the specific standards within each level of the rubric.

Crafting Scoring Rubrics: The Bottom-Up Approach With the bottom-up approach you begin with samples of students' work, using actual responses to create your own framework. Use examples of different quality levels to help you identify the dimensions along which students can be assessed. The following steps are helpful:

Step 1. Obtain copies of about 10 to 12 students' actual responses to a performance item. Be sure the responses you select illustrate various levels of quality of the general achievement you are assessing (e.g., science understanding, letter writing, critical reasoning, etc.).

Step 2. Read the responses and sort all of them into three groups: high-quality responses, medium-quality responses, and low-quality responses. Alternatively, you can ask students to do this. For tasks with which they have some experience (e.g., writing), and for which they therefore have some basis to begin to judge quality, this is a particularly powerful learning experience. The resulting bottom-up rubrics that students have helped create can be used for student self-evaluation and teacher-provided formative feedback.

Step 3. After sorting, carefully study each student's responses within the groups, and write (or have students write) very specific reasons why you put that response into that particular group. How are the students' responses in one group (e.g., high-quality group) different from the responses in each of the other groups? Be as specific as you can. For example, don't say they write better or have better ideas. Rather, say the students' sentences are more complex, or the students express unusual ideas in a very clear way. Write a specific and complete explanation on every student's response as to why it is placed into the group. Move a student's response into a different category if it turns out to fit better there.

Step 4. Look at your comments across all categories and identify (or have students identify) the emerging dimensions. In essence, you are creating your own conceptual framework in this step of the process. For example, if the responses are for a mathematics task, you may see computation, complete explanations, logical approach, and good mathematical reasoning as the dimensions.

Step 5. Separately for each of the three quality levels of each achievement dimension you identified in Step 4, write (or have students write) a specific student-centered description of what the responses at that level are typically like. You may have one to six achievement dimensions. The descriptions become the scoring rubric for marking new responses. Your final product may look similar to the top-down scoring rubric examples we presented previously, only having your own descriptions and dimensions.

The Northwest Regional Educational Laboratory (1998) has used the bottom-up approach extensively to train teachers to develop scoring rubrics. The two methods for creating rubrics do not necessarily lead to the same end product, and they are not equivalent procedures.

Validating Rubric Frameworks Adopted or Created Locally Compare the general (generic) rubric you draft with those developed by other districts, state assessment programs, national curriculum panels, or professional societies. Refine yours to make it clearer and more complete.

Creating a general rubric or a conceptual framework will help you maintain coherence and consistency in your scoring rubrics across all your

FIGURE 12.10 **Example of a holistic general scoring rubric for mathematics problem-solving tasks.**

Score level = 4

Mathematical knowledge

- Shows understanding of the problem's mathematical concepts and principles;
- Uses appropriate mathematical terminology and notations;
- Executes algorithms completely and correctly.

Strategic knowledge

- May use relevant outside information of a formal or informal nature;
- Identifies all the important elements of the problem and shows understanding of the relationships between them;
- Reflects an appropriate and systematic strategy for solving the problem;
- Gives clear evidence of a solution process, and solution process is complete and systematic.

Communication

- Gives a complete response with a clear, unambiguous explanation and/or description;
- May include an appropriate and complete diagram;
- Communicates effectively to the identified audience;
- Presents strong supporting arguments which are logically sound and complete;
- May include examples and counter examples.

Score level = 3

Mathematical knowledge

- Shows nearly complete understanding of the problem's mathematical concepts and principles;
- Uses nearly correct mathematical terminology and notations;
- Executes algorithms completely. Computations are generally correct but may contain minor errors.

Strategic knowledge

- May use relevant outside information of a formal or informal nature;
- Identifies the most important elements of the problems and shows general understanding of the relationships between them;
- Gives clear evidence of a solution process. Solution process is complete or nearly complete, and systematic.

Communication

- Gives a fairly complete response with reasonably clear explanations or descriptions;
- May include a nearly complete, appropriate diagram;
- Generally communicates effectively to the identified audience;
- Presents supporting arguments which are logically sound but may contain some minor gaps.

Score level = 2

Mathematical knowledge

- Shows understanding of the problem's mathematical concepts and principles;
- May contain serious computational errors.

Strategic knowledge

- Identifies some important elements of the problems but shows only limited understanding of the relationships between them;
- Gives some evidence of a solution process, but solution process may be incomplete or somewhat unsystematic.

Communication

- Makes significant progress towards completion of the problem, but the explanation or description may be somewhat ambiguous or unclear;
- May include a diagram which is flawed or unclear;
- Communication may be somewhat vague or difficult to interpret;
- Argumentation may be incomplete or may be based on a logically unsound premise.

Score level = 1

Mathematical knowledge

- Shows very limited understanding of the problem's mathematical concepts and principles;
- May misuse or fail to use mathematical terms;
- May make major computational errors.

Strategic knowledge

- May attempt to use irrelevant outside information;
- Fails to identify important elements or places too much emphasis on unimportant elements;
- May reflect an inappropriate strategy for solving the problem;
- Gives incomplete evidence of a solution process; solution process may be missing, difficult to identify, or completely unsystematic.

Communication

- Has some satisfactory elements but may fail to complete or may omit significant parts of the problem; explanation or description may be missing or difficult to follow;
- May include a diagram which incorrectly represents the problem situation, or diagram may be unclear and difficult to interpret.

Score level = 0

Mathematical knowledge

- Shows no understanding of the problem's mathematical concepts and principles.

Strategic knowledge

- May attempt to use irrelevant outside information;
- Fails to indicate which elements of the problem are appropriate;
- Copies part of the problem, but without attempting a solution.

Communication

- Communicates ineffectively; Words do not reflect the problem;
- May include drawings which completely misrepresent the problem situation.

Source: From "The Conceptual Framework for the Development of a Mathematics Performance Assessment Instrument," by S. Lane, 1992, *Educational Measurement: Issues and Practice,* 12(2), p. 23. Copyright 1992 by the National Council on Measurement in Education. Reprinted by permission of Blackwell Publishing.

performance tasks. Coherence applies not only to your assessment but also to your teaching, and it helps your students understand the standards that their learning should meet.

Administer the performance task to your students, then apply the general or specific rubric that you developed. If you have difficulty rating your students' performance, you should reexamine your rubric to see where it is unclear. Often you will need to expand the descriptions of each quality level in the rubric to include an example or to describe an aspect of student performance you initially forgot to include.

Checklists

A checklist consists of a list of specific behaviors, characteristics, or activities and a place for marking whether each is present or absent. You may use a checklist for assessing procedures students use, products students produce, or behaviors students exhibit. Students may use checklists to evaluate their own performance.

A **procedure checklist** assesses whether students follow the appropriate steps in a process or procedure. For example, a checklist may assess whether students are able to use a microscope properly. The form represents both the presence or absence of each step and the sequence that a particular student used to perform the task. Sometimes the major flaw in a student's performance is the order in which he or she performs the steps. Recording the correct sequence and the student's sequence on the form will help you attend to this aspect of performance.

A **product checklist** focuses on the quality of the things students make. Products include drawings, constructed models, essays, and term papers. These checklists identify the parts or other properties a product is supposed to have. You then inspect each product, checking whether those properties are present.

A **behavior checklist** consists of a list of discrete behaviors related to a specific area of student performance. For example, you may wish to identify the particular difficulties a student is having in the phonological, semantic, and syntactic aspects of spoken language. The behavior checklist might have items such as "uses only simple sentence structure" or "responds without delay to questions."

Students use a **self-evaluation checklist** to review and evaluate their own work. You could use the checklist students complete as a basis for a student-teacher conference in which you discuss a student's progress. As an example, consider the situation in which a student produces a best works mathematics portfolio. To create this portfolio, a student has to complete mathematics tasks and decide which of these completed tasks she should include in the portfolio. A student must select six or seven completed tasks to put into the best works portfolio. A checklist can help the student evaluate each entry and decide what to put into the portfolio. It can also serve as a basis for discussing the entries with peers, parents, or teachers. Because the checklist focuses on portfolio entries, it focuses the student's attention on the portfolio scoring rubric. However, the checklist is phrased in simpler and less formal language than the scoring rubric used by teachers. An adaptation of this checklist is shown in Figure 12.11.

How to Create Checklists

To create a checklist, you need a thorough understanding of the subject matter as well as the procedure or the product you want to assess. Without this knowledge you will find it difficult to identify critical performance and steps, critical flaws in the product, and potential student errors. To create checklists, complete a detailed analysis of the procedure you are evaluating or a careful specification of the precise characteristics of the desired student product.

Before crafting a *product checklist,* you should examine several students' products—especially those products that differ greatly in quality. Careful study of these products will help you identify the characteristics and flaws you want to include in your checklist.

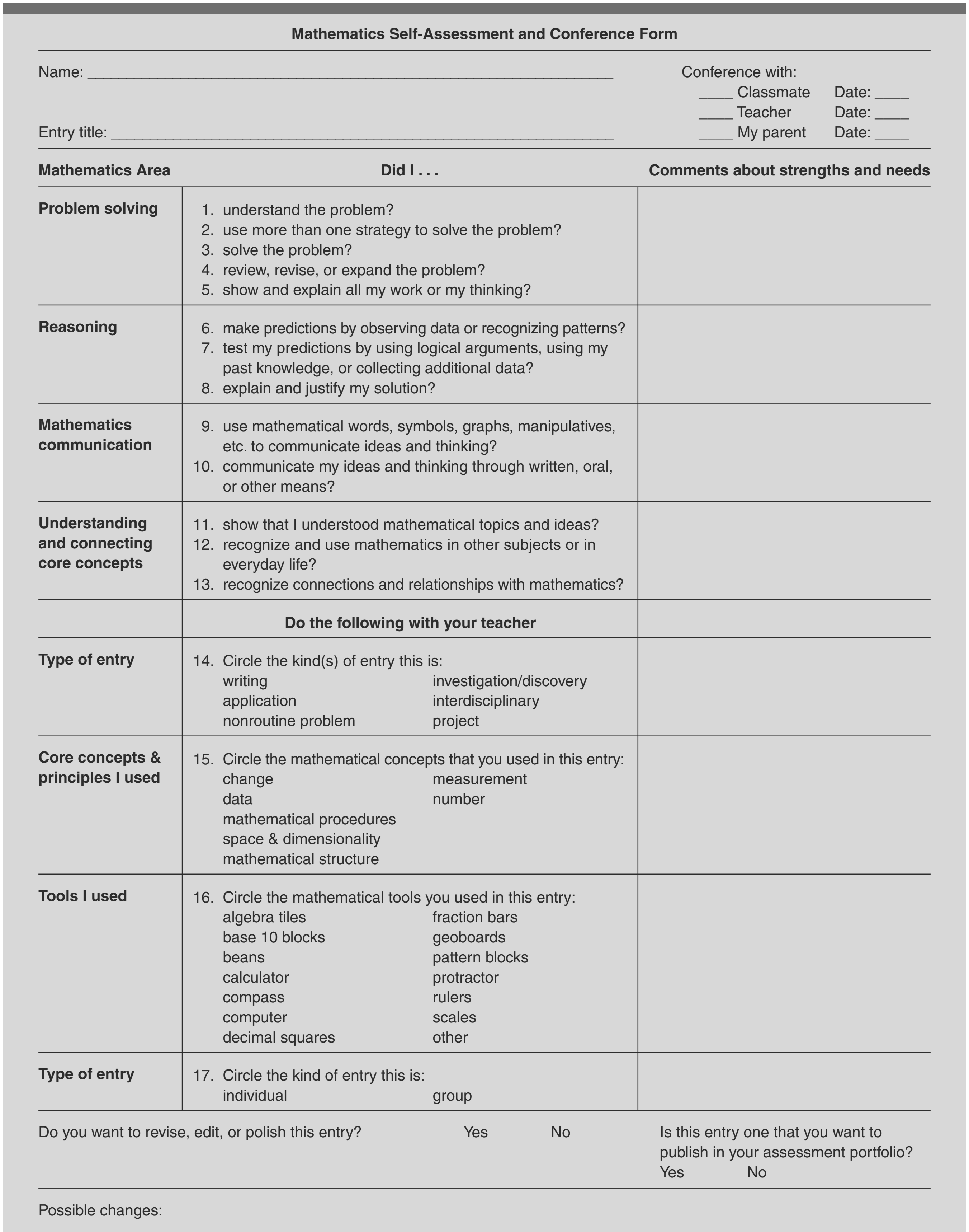
When crafting a *procedure checklist,* first observe and study students performing so you can identify all the appropriate steps. List each specific step in the procedure you want students to follow. Add to the list specific errors that students commonly make (avoid unwieldy lists, however). Order the correct steps and the errors in the approximate sequence in which they should occur. Note that if several equally correct procedures for accomplishing the learning target are available, developing a checklist this way will not be useful.

Rating Scales

Why Rating Scales Are Useful

Checklists help you evaluate whether a given step, a specific

FIGURE 12.11 **Example of a checklist that students use to evaluate their own entries for a best works portfolio.**



Mathematics Self-Assessment and Conference Form

Name: ____

Entry title: ____

Conference with:
____ Classmate Date: ____
____ Teacher Date: ____
____ My parent Date: ____

Mathematics Area	**Did I . . .**	**Comments about strengths and needs**
Problem solving	1. understand the problem? 2. use more than one strategy to solve the problem? 3. solve the problem? 4. review, revise, or expand the problem? 5. show and explain all my work or my thinking?	
Reasoning	6. make predictions by observing data or recognizing patterns? 7. test my predictions by using logical arguments, using my past knowledge, or collecting additional data? 8. explain and justify my solution?	
Mathematics communication	9. use mathematical words, symbols, graphs, manipulatives, etc. to communicate ideas and thinking? 10. communicate my ideas and thinking through written, oral, or other means?	
Understanding and connecting core concepts	11. show that I understood mathematical topics and ideas? 12. recognize and use mathematics in other subjects or in everyday life? 13. recognize connections and relationships with mathematics?	
	Do the following with your teacher	
Type of entry	14. Circle the kind(s) of entry this is: writing investigation/discovery application interdisciplinary nonroutine problem project	
Core concepts & principles I used	15. Circle the mathematical concepts that you used in this entry: change measurement data number mathematical procedures space & dimensionality mathematical structure	
Tools I used	16. Circle the mathematical tools you used in this entry: algebra tiles fraction bars base 10 blocks geoboards beans pattern blocks calculator protractor compass rulers computer scales decimal squares other	
Type of entry	17. Circle the kind of entry this is: individual group	

Do you want to revise, edit, or polish this entry? Yes No

Is this entry one that you want to publish in your assessment portfolio?
Yes No

Possible changes:

Source: Adapted from *Teacher's Guide: Kentucky Mathematics Portfolio,* by Kentucky Department of Education, 1993, Frankfort: Office of Assessment and Accountability, Author. Reprinted by permission.

property, or particular action is present. Many times you are concerned with more than the presence or absence of these elements. A rating scale assesses the *degree to which* students have attained the achievement dimensions in the performance task. As an example, consider assessing the quality of a student's oral presentation to the class. You would probably identify several dimensions of a "good oral presentation" and then judge the degree to which a student demonstrates each of them. A good oral presentation might include such characteristics as the degree to which a student presents material relevant to the topic; speaks in a smooth, unhesitating manner; uses correct grammar and language patterns; and makes visual contact with the audience. You need to assess and record the degree to which a student demonstrates each dimension, rather than assessing on an all-or-none, present-or-absent basis.

Rating scales can be used for teaching purposes as well as assessment:

1. *The rating scale helps students understand the learning target and focus their attention on the important aspects of the performance.* You can give it to students as they prepare for the performance task.
2. *The completed rating scale gives specific feedback to students concerning the strengths and weaknesses of the performance.* You can give the rating scale to students after you have used it to evaluate their performance.
3. *Students not only achieve the learning targets but also may internalize the criteria used to evaluate their achievement.* This will help them automatically apply the criteria in the rating scales to their work. To accomplish this, you must rate the same achievement dimensions across several different performance tasks throughout the year.
4. *Ratings help you show students individual growth.* If you keep copies of ratings in a file, you will have a record to help you monitor and assess each student's progress. To do this effectively, you need to use the same (or similar) rating scale across all tasks. To ensure that the information you collect is comparable from occasion to occasion, use a general rubric or a framework to create specific rubrics and rating scales.

Types of Rating Scales Although there are many varieties of rating scales, three varieties—numerical rating scales, graphic rating scales, and descriptive graphic rating scales—when used to their full advantage, serve the teacher well for most purposes.

To use a **numerical rating scale**, you must mentally translate judgments of quality or degree of achievement into numbers. Figure 12.12 shows one example of this approach. The teacher of a technical drawing course lists 10 achievement dimensions against which he evaluates each drawing. He rates students' achievement of each dimension on a scale of 0 to 10 and then adds up the ratings. If a particular dimension—for example, "quality of arcs, circles, and tangents"—does not apply to a particular kind of drawing, then it is omitted. Figure 12.12 also shows the results of using the rating scale to evaluate a ninth grader's drawing.

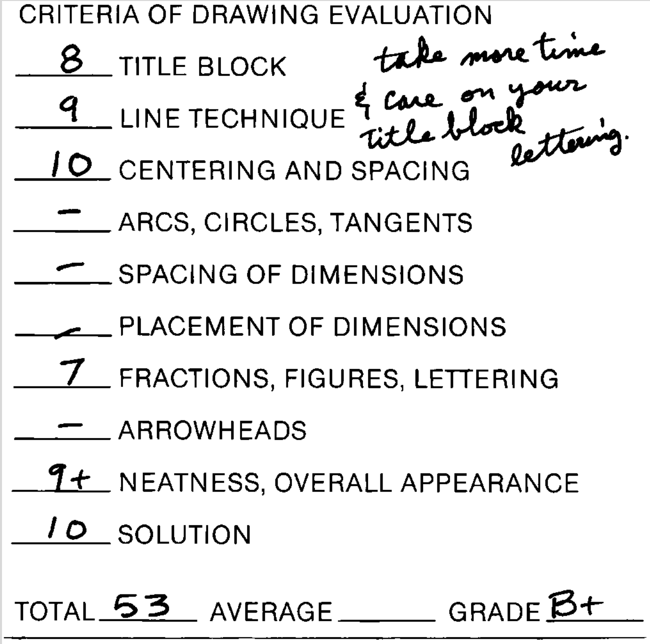
Notice from the example that simply providing students with "numbers" is not sufficient. You need to make verbal comments—both positive and negative—to give students the feedback necessary to make improvements. In addition, you may give students the list of criteria and ask them to edit their own work before turning in their assignments.

You will increase objectivity and consistency in results from numerical rating scales if you provide a short verbal description of the quality level each number represents. Alternately, you can associate each numerical level with an example or actual specimen of the products you are rating. You can match a student's performance to either a verbal description or an actual specimen assigning the corresponding number. The *Thorndike Handwriting Scale* illustrated in Figure 12.13 is an example of the latter.

Graphic rating scales use an unbroken line to represent the particular achievement dimension on which you rate a student's performance or product. Verbal labels describing levels of quality define different parts of the line. This guides you in deciding which ratings to assign to a student. Figure 12.14 is an example of a simple graphic rating scale that a teacher might use to rate a student's attainment of cooperative learning targets in a group project. In Figure 12.14, the endpoints of the line are "anchored" by Never and Always; Seldom, Occasionally, and Frequently define intermediate levels of achievement.

On a graphic rating scale, you can check any point along the line, not just the defined points. Thus, the graphic rating scale does not force your rating into a discrete category or into being a whole number, as does the numerical rating method. In practice, a serious problem with the use of verbal labels such as *usually, seldom,* and *frequently* is that

FIGURE 12.12 Example of an analytic rubric in the form of a numerical rating scale used to assess a student's technical drawing (i.e., product assessment).



CRITERIA OF DRAWING EVALUATION

8 TITLE BLOCK

9 LINE TECHNIQUE

10 CENTERING AND SPACING

– ARCS, CIRCLES, TANGENTS

– SPACING OF DIMENSIONS

– PLACEMENT OF DIMENSIONS

7 FRACTIONS, FIGURES, LETTERING

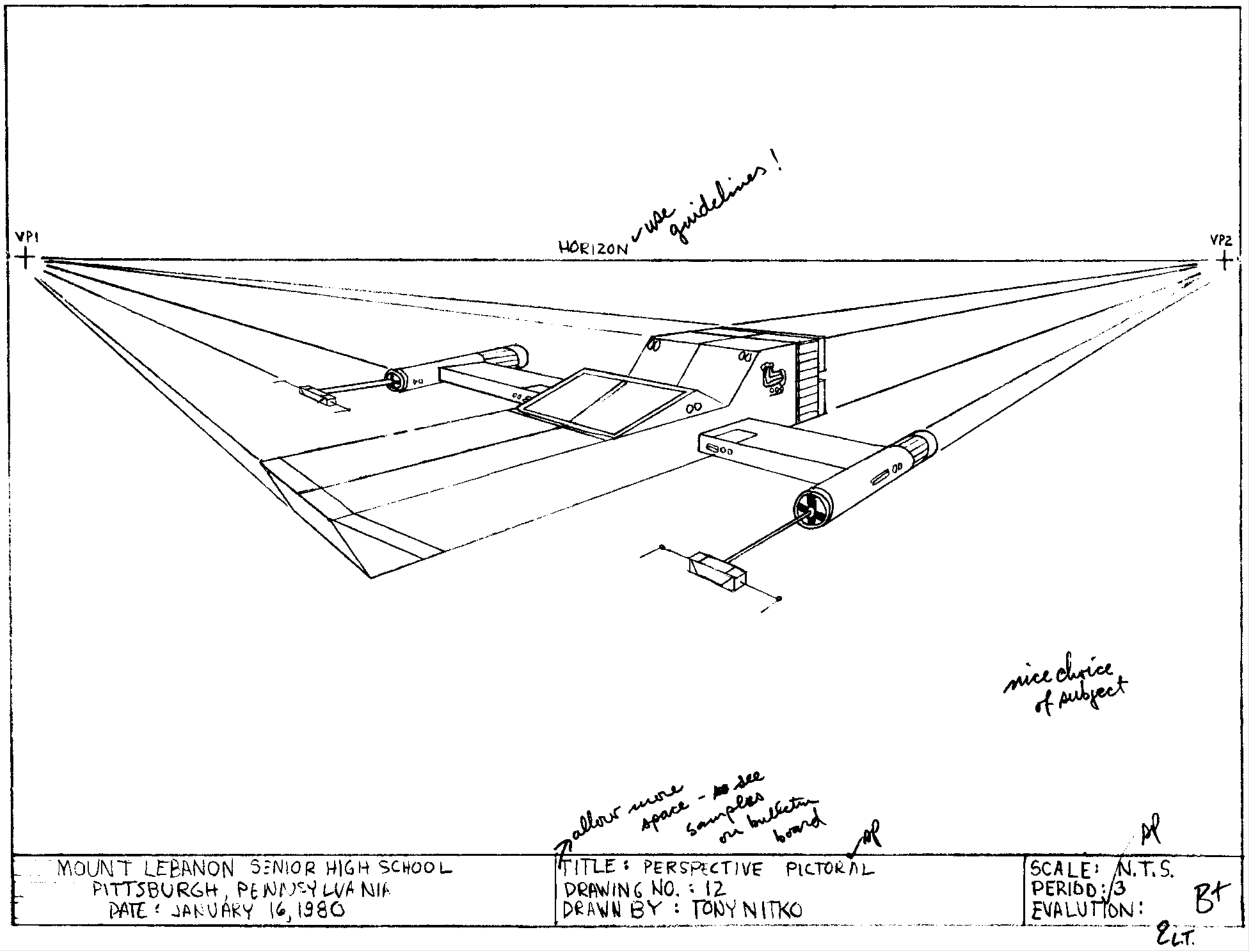
– ARROWHEADS

9+ NEATNESS, OVERALL APPEARANCE

10 SOLUTION

TOTAL 53 AVERAGE ______ GRADE B+

take more time & care on your title block lettering.

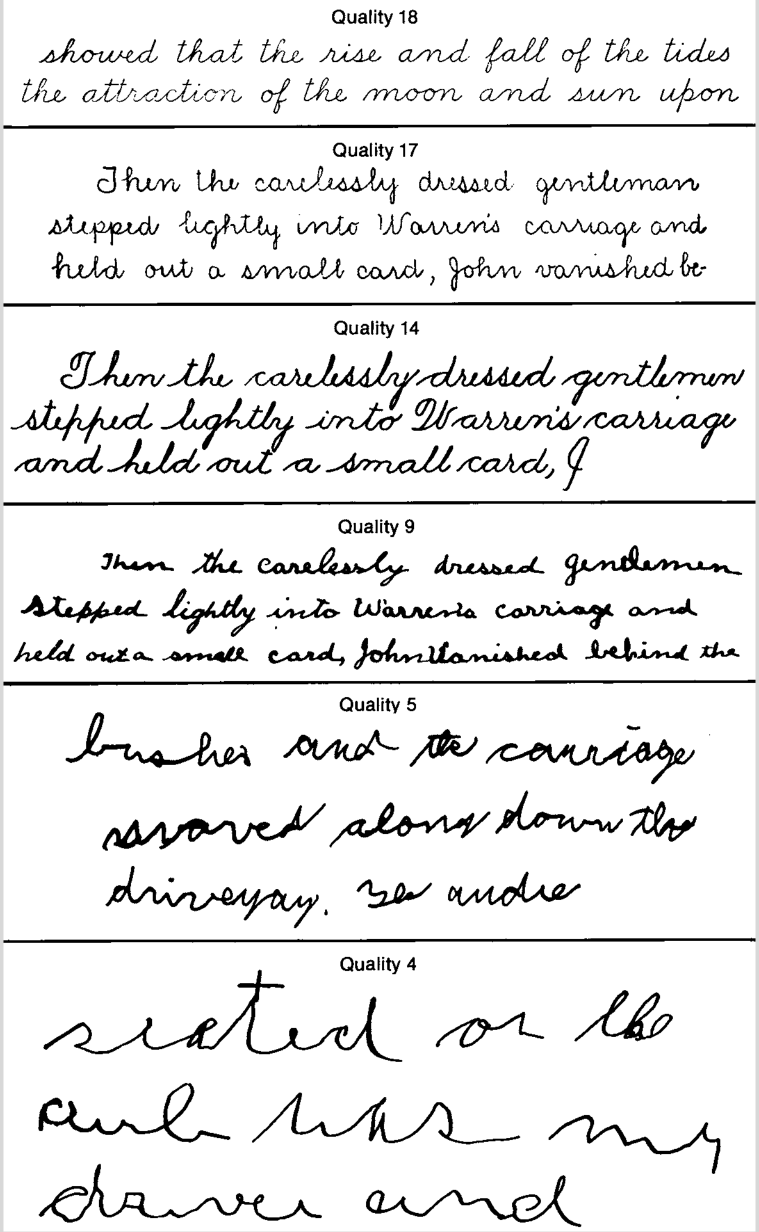


MOUNT LEBANON SENIOR HIGH SCHOOL PITTSBURGH, PENNSYLVANIA DATE: JANUARY 16, 1980	TITLE: PERSPECTIVE PICTORAL DRAWING NO.: 12 DRAWN BY: TONY NITKO	SCALE: N.T.S. PERIDD: 3 EVALUTION: B+

2 L.T.

Source: Rating scale by permission of Mr. Scott Patton, technical drawing instructor at Mt. Lebanon High School; drawing courtesy of Anthony Nitko Jr.

FIGURE 12.13 Example of a scale (Thorndike's) for measuring handwriting. A series of handwriting specimens were scaled on a numerical "quality" scale. To use the scale a student's sample of writing is matched to the quality of one of the specimens and assigned the given numerical value. This figure shows only some of the specimens.



Quality 18

showed that the rise and fall of the tides
the attraction of the moon and sun upon

Quality 17

Then the carelessly dressed gentleman
stepped lightly into Warren's carriage and
held out a small card, John vanished be-

Quality 14

Then the carelessly dressed gentlemen
stepped lightly into Warren's carriage
and held out a small card, J

Quality 9

Then the carelessly dressed gentlemen
Stepped lightly into Warren's carriage and
held out a small card, John vanished behind the

Quality 5

bushes and the carriage
swayed along down the
driveyay. The audie

Quality 4

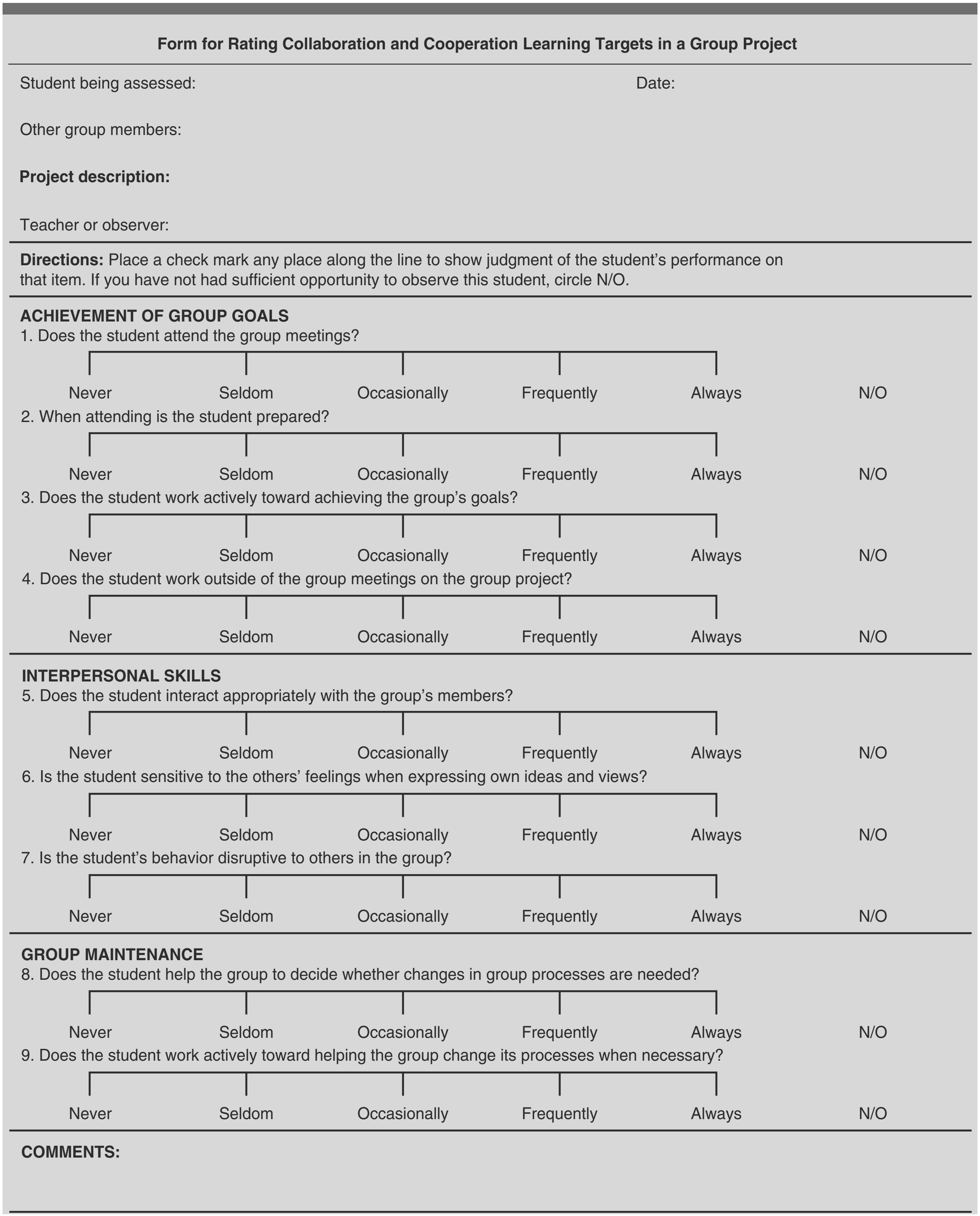
seated on the
curb like my
drawer and

Source: From "Handwriting," by E. L. Thorndike, 1910, *Teacher's College Record, (11)* (2), pp. 62, 65, 70, 73.

they are undefined; different raters do not agree on what they mean. Defining the levels on the scale with more behavioral descriptions makes your ratings much more consistent and meaningful.

A **descriptive graphic rating scale** is a better format for rating. This type of scale replaces the ambiguous single words (e.g., *frequently*) with short behavioral descriptions of the various points along the scale. Each degree of success on each dimension is defined by a brief description. Sometimes numbers are also printed along the line, combining the features of both a numerical and a graphic rating scale. Describing the points of the scale by behavior descriptions leads to increased consistency of ratings across raters and students. One example of a descriptive graphic rating scales is the scale for rating a student's disposition toward critical thinking (Figure 11.10) in Chapter 11. Figure 12.15 presents

FIGURE 12.14 **A simple graphic scale for assessing cooperative learning targets with a group project.**



Form for Rating Collaboration and Cooperation Learning Targets in a Group Project

Student being assessed: Date:

Other group members:

Project description:

Teacher or observer:

Directions: Place a check mark any place along the line to show judgment of the student's performance on that item. If you have not had sufficient opportunity to observe this student, circle N/O.

ACHIEVEMENT OF GROUP GOALS

1. Does the student attend the group meetings?

Never Seldom Occasionally Frequently Always N/O

2. When attending is the student prepared?

Never Seldom Occasionally Frequently Always N/O

3. Does the student work actively toward achieving the group's goals?

Never Seldom Occasionally Frequently Always N/O

4. Does the student work outside of the group meetings on the group project?

Never Seldom Occasionally Frequently Always N/O

INTERPERSONAL SKILLS

5. Does the student interact appropriately with the group's members?

Never Seldom Occasionally Frequently Always N/O

6. Is the student sensitive to the others' feelings when expressing own ideas and views?

Never Seldom Occasionally Frequently Always N/O

7. Is the student's behavior disruptive to others in the group?

Never Seldom Occasionally Frequently Always N/O

GROUP MAINTENANCE

8. Does the student help the group to decide whether changes in group processes are needed?

Never Seldom Occasionally Frequently Always N/O

9. Does the student work actively toward helping the group change its processes when necessary?

Never Seldom Occasionally Frequently Always N/O

COMMENTS:

FIGURE 12.15 Example of a simple rating scale to use as a scoring rubric for assessing the quality of a student's oral or written presentation of an argument.

Student's name: Date:
Topic:

1. Did the student clearly state the thesis or main point?

0	1	2	3
Did not state or imply the main point or thesis	Implied the main point or thesis but did not state it clearly	Stated the main idea or thesis clearly but only matter-of-factly	Stated the main idea or thesis clearly, enthusiastically, and interestingly for the audience

2. Did the student define the key terms when necessary to do so?

0	1	2	NA
No attempt to define key terms, even when it was necessary to do so	Attempts to define the key terms, but was not effective in doing so	Clearly and effectively defines the necessary key terms	The presentation was such that defining key terms was unnecessary

3. Did the student use sound reasoning to support the main point or thesis?

0	1	2	3
Offered no supporting reasons for the thesis or position taken	Supporting reasons given but they are off-target or they do not lend direct support for the thesis	Gives relevant supporting reasons, but could have given better or more diverse reasons	Gives excellent supporting reasons, good diversity, directly applicable

4. Did the student use relevant facts in appropriate ways to support the thesis?

0	1	2	3
Gave no facts, used completely irrelevant facts, or cited facts from noncredible sources	Gave facts to support the thesis but the generalizations from them were weak, somewhat inappropriate, or incomplete; facts are cited from credible sources	Gives several appropriate facts that support the thesis, generalizations from facts are appropriate, sources for facts are credible	Gives highly appropriate facts, excellent generalizations from facts that support the thesis, sources of facts are credible, facts used well in making the argument

5. Did the student portray and evaluate alternative positions fairly?

0	1	2	3
Alternative positions not mentioned and not evaluated	Alternative positions are mentioned but they are either not portrayed fairly, not evaluated properly, or not relevant to the thesis	Some of the relevant alternative positions are mentioned, they are portrayed properly, and evaluated properly; other important alternative positions are omitted	All relevant and important alternative positions are mentioned, presented fairly, and evaluated properly

6. Did the student rebut the alternative positions well?

0	1	2	3
No attempt was made at rebuttal	Attempts at rebuttal were made but they are ineffective or incomplete	Rebutted adequately, but could have been more effective in explaining the shortcomings of the alternatives	Rebutted well, was effective, clear about the inadequacies of the alternatives, convincingly presented

7. Did the student present a well-organized argument?

0	1	2	3
Organization was disconnected, lacked direction, confused the thesis or main point	Organization was clear, but not effective; connections to thesis or main point were not sharp; details were often out of place	Organization was good and contributed to the effectiveness of the argument, but a few details were out of place; sometimes the connections to the main point were weak or out of place	Organization was very clear and enhanced the argument; the presentation kept the audience interest focused on the main issues

another example, for rating a student's oral or written presentation of an argument.

Rating Scale Errors and How to Avoid Them

Several common errors occur when teachers rate students. Teachers who do not use all of the points on a rating scale cause the following errors:

- **Leniency error** occurs when a teacher tends to make almost all ratings toward the high end of the scale, avoiding the low end.
- **Severity error** is the opposite of leniency error: A teacher tends to make almost all ratings toward the low end of the scale.
- **Central tendency error** occurs when a teacher hesitates to use extremes and uses only the middle part of the scale. Central tendency errors sometimes occur when a teacher has to make strong inferences about students (e.g., regarding "creativity" or "dedication") and, in hesitation, the teacher tends to mark nearly everyone as average. Central tendency errors may occur when a teacher does not know the students very well.

Using only certain parts of the rating scale has two negative consequences. First, when you give only very high, very low, or "middle" ratings, you introduce your own quirks and biases into the ratings, thus lowering their validity for describing students' ability in performing the task. Second, when your ratings bunch up and do not distinguish one student's performance from another's, they become unreliable, which in turn reduces the validity of the scores.

There are other common rating scale errors. We mentioned these errors in Chapter 10, because they happen during essay scoring, as well.

- A **halo effect** occurs when teachers lets their general impressions of students affect how they rate the students on specific dimensions. For example, if you gave a student a higher rating for his project than the student deserves because you "just know" that the student is "really" very good, you would be committing the halo effect error.

  The general "halo" you place around the student affects your ability to judge the student's standing on specific performances. (The halo effect may *work in reverse,* of course: Your general impression of a student as "not very good" may lead you to lower ratings on specific dimensions more than the student deserves.) One expression of the halo effect may occur when teachers need to make grading decisions for students whose assessment results put them on the border between two letter-grade categories: The error is that individuals who favorably impress a teacher are moved into the upper category; those who less favorably impress the teacher are moved into the lower category.
- **Personal bias** occurs when teachers tend to rate based on inappropriate or irrelevant stereotypes favoring boys over girls, whites over blacks, working families over welfare recipients, or particular families and individual students the teachers may dislike.
- A **logical error** occurs when teachers give similar ratings on two or more dimensions of performance that the teachers believe are logically related but that are in fact unrelated. For example, a teacher may falsely believe that students with exceptionally high scores on scholastic aptitude tests also should be the top students in all subject areas. The teacher then marks the high-scoring aptitude test students differently from the way the low scorers are marked.

  Logical errors are a result of a teacher's ignorance and unfounded beliefs, rather than the teacher's personal quirks and biases about individuals or groups of students.

Other errors occur when "outsiders" rate performance assessments. When states and large school districts implement performance assessments, individuals other than their teachers usually rate students' work. In these cases, the raters are trained in and practice using a particular scoring rubric.

- **Rater drift** occurs when the raters, whose ratings originally agreed, begin to redefine the rubrics for themselves. As a result, the raters no longer produce ratings that agree with the original rubrics even though they were trained on the same rubrics. The remedy for this is to monitor the ratings and to retrain those raters who appear to have drifted away from agreed-on standards.
- **Reliability decay** is a related error: Immediately after training, raters apply the rubrics consistently across students and mark consistently with one another. However, as time passes, the ratings become less consistent, both across students and across raters. Monitoring and retraining are remedies for this effect, too.

Evaluating Scoring Rubrics and Rating Scales

The checklist below provides guidance for evaluating scoring rubrics and other classroom rating scales. You will use a specific scoring rubric to assess students' performance on a particular performance task; we have already provided a checklist for evaluating your performance task. You need to evaluate the scoring rubric, too.

✔ **Checklist**

A Checklist for Judging the Quality of Scoring Rubrics and Rating Scales

Ask these questions of every rubric or rating scale you write. If you answer no to one or more questions, revise the rubric or rating scale accordingly.

1. Overall, does the rubric emphasize the most important content and processes of the learning targets?
2. Will the scores you get from the parts of your rubric (i.e., achievement dimensions) match the emphasis you gave them in your assessment plan?
3. Does the maximum possible total number of marks (points) obtained from the rubric match the emphasis you gave these learning targets in your assessment plan?
4. Will your students understand the rubric?
5. Are the categories rated with the rubric suitable for giving students the guidance they need to improve their performance on the learning targets?
6. Is the rubric for this particular task a faithful application of the general rubric or conceptual framework?
7. Are the levels for the scales for the parts of the rubric (i.e., the levels of the achievement dimensions) described clearly in terms of performance you can observe students doing?
8. With regard to this particular task, does the rubric allow you to assess the students' use of the appropriate:
   a. declarative and procedural content dimensions?
   b. processes that are important to the learning target(s)?
9. If the purpose of this task is to assess students' use of alternative correct answers/products or alternative correct processes/strategies, does the rubric clearly describe how each is to be rated and marked?
10. Does the rubric allow you to distinguish a wide range of students' achievement levels on this task, rather than putting all students into one or two achievement levels?

How to Make Rubrics and Rating Scales Consistent With Your Grading System

We have already mentioned that your letter-grade system, your scoring rubrics, and your rating scales should have the *same number of levels*. For example, if your grading system has A, B, C, D, or F, the rubrics should have five levels, not four or six. When you write your scale level descriptions, make the meaning of these levels consistent with the meaning of grades, too. Consistencies in scales across all assessments are important ways to improve the validity of your assessment results. It will be difficult for you to report student progress on report cards with letter grades if your scoring rubrics and rating scales are not aligned to the grading system.

Improving Reliability of Rubrics and Ratings

The **reliability of ratings** is an important criterion for evaluating performance assessments. Many of the suggestions for improving the reliability of grading essays (Figure 10.6) apply to performance assessments, as well. The following reliability coefficients (discussed in Chapter 4) are among those appropriate to use with the more continuous scores awarded to students from performance assessments. For classroom performance assessments, scorer reliability (teacher consistency in marking) and alternate forms (differences due to students selecting a different tasks, if that was allowed) are usually the most problematic.

Estimating Reliability Over Time

- Test-retest
- Alternate forms on different occasions

Estimating Reliability on a Single Occasion

- Alternate forms
- Coefficient alpha
- Split-halves coefficient

Estimating Scorer Reliability

- Correlation of two scorers' results
- Percentage of agreement
- Kappa coefficient

Appendix J shows how to calculate coefficient alpha, split-halves percentage agreement, and kappa coefficient. However, the appendix shows percent agreement and kappa calculations only for the special case in which pass-fail or mastery-nonmastery decisions are made. Although these two indices can easily be applied to scores in more than two categories, that is beyond the scope of this book.

Figure 4.8 in Chapter 4 discussed how to improve the reliability of assessment results. Here are additional suggestions that apply specifically to improving the reliability of ratings from scoring rubrics and rating scales:

1. Organize the achievement dimensions within a scoring rubric into logical groups that match the content and process framework of the curriculum.
2. For each achievement dimension, use behavioral descriptors to define each level of performance.
3. Provide specimens or examples of students' work to help define each level of an achievement dimension.
4. Have several teachers work together to develop a scoring rubric or rating scale.
5. Have several teachers review and critique the draft of a scoring rubric or rating scale.
6. Provide training and supervised practice for all persons who will use the scoring rubric or rating scale.
7. Have more than one rater rate each student's performance on the task.
8. Monitor raters by periodically sampling their ratings, checking on the accuracy and consistency with which they are applying the scoring rubrics and rating scales. Retrain those persons whose ratings are inaccurate or inconsistent.

DESIGNING PROJECTS

Projects are usually worthwhile educational activities. Their usefulness as performance assessment tasks for individual students, however, depends on four conditions. You must ensure that (1) you and your students are very clear that the project focuses on one or more important curriculum learning targets, (2) each student does his or her own work, (3) each student has equal access to the resources needed to prepare an excellent final product and to achieve an excellent evaluation, and (4) you can control your own biases toward certain types of products and fairly evaluate other well-done projects. Middle- and upper-middle-class students with highly educated parents, for example, often have access to more resources than their less fortunate peers. You may tend to evaluate such students' projects more highly because they use these resources and produce very good-looking products. However, by so doing you may be biasing your evaluations toward certain social classes of students.

You can overcome the previously mentioned limitations by carefully planning for using projects *as classroom activities and educational assessment tasks,* rather than only as *classroom activities*. You do this by designing projects with the criteria for designing performance assessments in mind. Following these suggestions should help you craft useful projects for assessment purposes:

1. Explicitly define the most important learning targets for which the project will provide you with a direct assessment opportunity.
2. Identify specific characteristics and achievement dimensions of the final project that are most strongly linked to the learning targets that you are evaluating. Evaluate students only on these dimensions.
3. Define a continuum of levels of achievement for each dimension. Use each student's location on this scale as your assessment.
4. Design the scoring rubric you will use for evaluating each achievement dimension you will assess for the project.
5. Define the weight you will give the marks from each achievement dimension when you calculate the overall project grade.
6. Limit the resources students may use to complete the project if students vary widely in their ability to access resources.

How to Manage Your Classroom Projects

Because projects usually span several weeks, you must plan to manage them. Monitor individual students to be sure they are making regular progress. Mentor students to help them overcome operational problems that may be beyond their control (e.g., a key person students were to interview for the project has become ill and cannot see them). Mentor students to keep them focused on completing the project. Monitor the procedures and processes the students are using to ensure they will be able to address the learning targets set for the project.

The following classroom project management strategies can help meet these project management goals:

Strategy 1. *Clarify the outcome(s) you expect.* Be sure each student thoroughly understands both the purpose(s) of the project (the learning target being assessed) and what you expect the project to look like. Show and discuss examples of high-quality projects you saved from former students.

Strategy 2. *Put your expectations in writing.* Distribute to and discuss with students a written description of what you expect in the way of a project, processes, and the major purpose of the project.

Strategy 3. *Clarify the standards you will use to evaluate the project.* Explain and give students copies of the scoring rubrics you will be using to evaluate the project.

Strategy 4. *Let students participate in setting standards.* Each student should internalize the quality standards and have a sense of ownership of them. Use past projects to help students induce achievement dimensions. Help students describe the quality levels within each dimension.

Strategy 5. *Clarify deadlines.* Set deadlines that are long enough so students can develop and complete authentic projects. However, make the time frame short enough to be practical and so that students must keep on task, have no time to waste, and finish on time.

Strategy 6. *Require progress reports.* For longer projects, specify weekly or biweekly dates for students to report their progress (e.g., every Friday). This helps keep the students on task and allows you to assess the processes they use and their progress toward completing the project. It also alerts you to any problems beyond student control that may require your intervention. Use these reports as opportunities to give formative feedback to all students.

Strategy 7. *Minimize plagiarism opportunities.* Students should work to the best of their ability. Explain to students what constitutes plagiarism and the importance of doing one's own work even though it is not perfect. Avoid projects that may inadvertently encourage students to plagiarize material. Projects that help reduce the students' temptation to copy include interviewing, comparing opinions, making models, and the like.

DESIGNING PORTFOLIOS

Although some consider portfolios a type of performance assessment, they differ enough from other types that we will discuss them separately. For purposes of assessment, a **portfolio** is a limited collection of a student's work used either to present the student's best work(s) or to demonstrate the student's educational growth over a given time. A portfolio is not simply a scrapbook or collection of all of a student's work. The works put into a portfolio are carefully and deliberately selected so the collection as a whole accomplishes its purpose. Many authors have lists of different types of portfolios. Most of them fall into one of two assessment purposes: presenting one's best work or demonstrating educational growth.

A best works portfolio contains a student's best final products. You use best works portfolios primarily for summative purposes. Here are examples of some of the purposes that best works portfolios serve:

Examples

Examples of Best Works Portfolios

General purpose: Evaluation of individual students

Possible specific purposes

- Evidence of subject-matter mastery and learning.
- Evidence of high-level accomplishment in an area such as art or writing.

- Evidence of minimal competence in a subject for purposes of graduation.
- Evidence of a school district's accomplishments.

Contents of the portfolio

- A student's best works are selected to provide convincing evidence that the student has achieved specific learning targets.

General purpose: Communications

Possible specific purposes

- A student's showcase for his or her parents.
- Pass on information about a student to the next teacher.
- A school's showcase.

Contents of the portfolio

- Examples of accomplishments that may be typical or may impress others.

Very often the contents of the best works portfolio are prescribed. For example, to certify a student's accomplishment in art, educational authorities may require a drawing, a painting, a sculpture, a craft product, and one work in a medium of the student's choosing. In mathematics, an educational authority may require that a student's portfolio contain a table of contents, a letter telling the portfolio evaluator about the entries included, and five to seven best works involving a variety of types of activities, tools, and topics (Kentucky Department of Education, 1993a). Scoring rubrics for portfolios usually apply to the entire portfolio rather than to each piece separately, but there are exceptions.

Students need to learn how to create a best works portfolio to present themselves in the best possible way. Among the portfolio-making skills students need to learn are deciding exactly what they want to communicate or accomplish through the portfolio; how to choose the pieces to include in the portfolio; how best to present the pieces chosen; and evaluating the qualities of the pieces selected using the scoring rubrics that will be applied to their portfolios.

A growth portfolio contains examples of a student's work, along with comments, that demonstrate how well the student's learning has progressed over a given period. It does not focus on the final products a student produces. Instead, you and the student use the portfolio for formative purposes to monitor the student's learning and thinking progress, to diagnose learning and thinking difficulties, and to guide new learning and thinking. The student plays a significant role in deciding what should be included in this portfolio and learns to use the portfolio to understand and evaluate her own progress. Here are some examples:

Examples

Examples of Growth Portfolios

General purpose: Monitoring progress of individual students

Possible specific purposes

- Teachers and/or students want to review progress and change in achievement.
- Student needs to look over his or her work to see the "long view" or "whole picture" of what has been accomplished.

Contents of the portfolio

- A student's products or works that appear at intermediate stages in the course of the student's learning. These may include early drafts, records of thinking, and rewrites. The final product is placed into the portfolio, too.

General purpose: Daily instruction

Possible specific purposes

- A basis for discussing with the student individual ideas and work.
- Keep a record of changes in a student's thinking and conceptual explanations.
- A basis for diagnosing a student's learning difficulties in a subject.

Contents of the portfolio

- Examples of a student's recently completed work, data the student collected, recent findings from an ongoing investigation in the subject matter, a student's own explanations of the work that is under way, and so on.

The clearer you are about your portfolio learning targets and purpose(s), the better you are able to design it. If the portfolio must serve more than one purpose, you will need to consider carefully the focus of each portfolio entry, so that each entry serves at least one of your intended purposes.

Growth Portfolio Organization and Contents

Initial Planning for a Growth Portfolio The purpose of a growth portfolio is to serve as a tool for you and students to monitor learning, diagnose difficulties, guide new learning, and show progress and development. For a growth portfolio to be effective, you must carefully design it using the following principles:

1. *Be very clear about the learning targets toward which you wish to monitor students' learning progress.* The clearer you are, the better your portfolio system will be.
2. *Have a firm understanding of a learning progress theory.* The theory you choose to follow will guide you to identify what you should look for when assessing changes in a student's conceptual development or in diagnosing a student's learning difficulties.
3. *If several teachers in a school are committed to using growth portfolios, collaborate and work cooperatively with them.* If teachers coordinate the general approach, contents to be included, and the portfolio organization, students will not be confused and will receive a consistent message about the nature and purposes of their portfolio activities.
4. *Use some type of rubric to define assessment criteria and to help you be consistent in how you apply these criteria, both across students and with the same student over time.*

Figure 12.16 is an example of how one school's language arts teachers planned and organized their students' growth portfolios. The organization was decided by a group of teachers working together. Time frames are spelled out, as are content and the person who should contribute each entry (both the student and the teacher contribute in this example). Notice that the teachers specified learning targets, and for each target they described the type of entries that should be included.

Students' Self-Evaluation Entries in Growth Portfolios Notice that the plan shown in Figure 12.17 calls for students to reflect on and evaluate their own progress as readers and writers. To facilitate students' ability to meet these types of learning targets, the teachers found it necessary to design special portfolio entry forms. These forms contained the questions students are to ask of themselves and a place for them to answer the questions. Without such organization, it would be easy for the portfolio to become too disorganized for the teachers to use efficiently in class. Figure 12.17 is an example of one of these entry forms completed by a student.

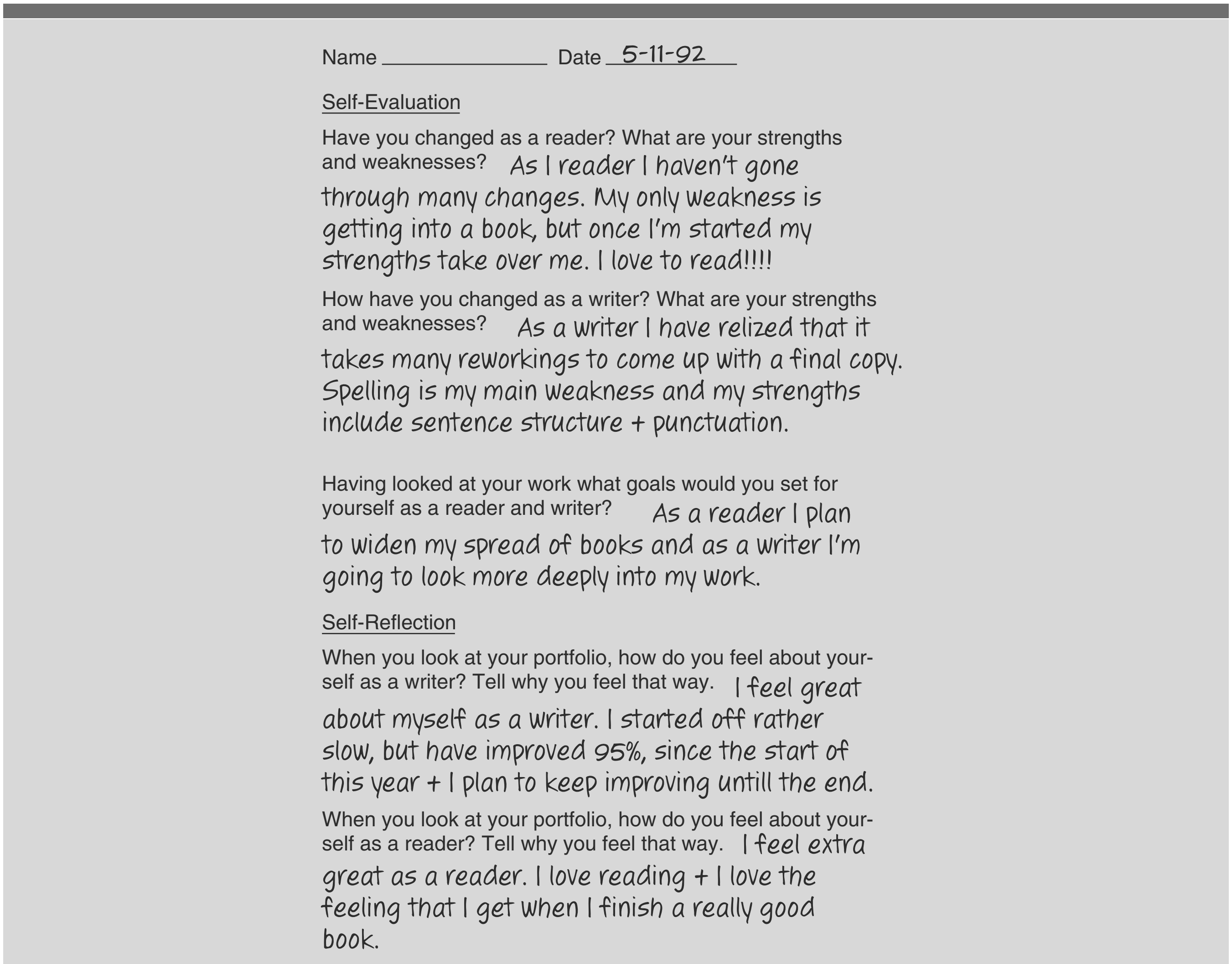
Using Growth Portfolios in Your Teaching Growth portfolios, like other assessments, work best when integrated fully into your teaching. Some

FIGURE 12.16 **Example of how the Bellevue, Washington, teachers organized their students' reading and writing portfolios to assess growth.**

Learning Target	What is put into the portfolio	Frequency of entry and assessment
1. Develop a meaningful ownership of one's own learning and work to be evaluated	1. (a) Student-selected pieces of work (b) Entry slip explaining why each piece was included	1. Three or more times per year
2. Evaluate one's own progress over time	2. (a) Student reviews his or her own portfolio (b) Student answers questions about his or her development as a reader and writer	2. Two or more times per year
3. Interact with the text to create meaning	3. Entry slip retelling the piece read or explaining its meaning	3. Two or three times per year
4. Choose to read a variety of material	4. Log of books/articles read during a two-week period	4. Two or three times per year
5. Communicate effectively through writing	5. Samples of longer pieces of writing	5. Two or three times per year
6. Student develops as a reader and writer	6. (a) Student drafts, notes, and other work selected by the teacher (b) Teacher's notes and comments about the student's progress	6. Left to the teacher's discretion

Source: Adapted from "Literacy Portfolios for Teaching, Learning, and Accountability: The Bellevue Literacy Assessment Project," by S. W. Valencia and N. A. Place, in *Authentic Reading Assessment: Practices and Possibilities* (pp. 139–141), by Sheila W. Valencia, E. H. Hiebert, and Peter P. Afferbach (Eds.). Copyright © 1994 by the International Reading Association. Adapted by permission.

FIGURE 12.17 **Example of a middle school student's portfolio entry after evaluating her own reading and writing progress over the year.**



Name ______________ Date 5-11-92

Self-Evaluation

Have you changed as a reader? What are your strengths and weaknesses? As I reader I haven't gone through many changes. My only weakness is getting into a book, but once I'm started my strengths take over me. I love to read!!!!

How have you changed as a writer? What are your strengths and weaknesses? As a writer I have relized that it takes many reworkings to come up with a final copy. Spelling is my main weakness and my strengths include sentence structure + punctuation.

Having looked at your work what goals would you set for yourself as a reader and writer? As a reader I plan to widen my spread of books and as a writer I'm going to look more deeply into my work.

Self-Reflection

When you look at your portfolio, how do you feel about yourself as a writer? Tell why you feel that way. I feel great about myself as a writer. I started off rather slow, but have improved 95%, since the start of this year + I plan to keep improving untill the end.

When you look at your portfolio, how do you feel about yourself as a reader? Tell why you feel that way. I feel extra great as a reader. I love reading + I love the feeling that I get when I finish a really good book.

Source: Adapted from "Literacy Portfolios for Teaching, Learning, and Accountability: The Bellevue Literacy Project," by S. W. Valencia and N. A. Place, in *Authentic Reading Assessment: Practice and Possibilities* (p. 146), by Sheila W. Valencia, E. H. Hiebert, and Peter P. Afferbach (Eds.). Copyright © 1994 by the International Reading Association. Adapted by permission.

writers advocate making the portfolio the center of your instructional planning and teaching activities so you and your students will interact intensively with the portfolio contents. This is called the **portfolio culture model** of conceptual change (Duschl & Gitomer, 1991; Niyogi, 1995). In a portfolio culture, instructional activities and projects are opportunities for students to record their intermediate progress, their progressive understanding of concepts and phenomena, and their interactions with peers and teachers. Duschl and Gitomer suggest the work included must have certain characteristics to be useful in a portfolio culture educational setting that focuses on restructuring students' conceptual development. The following suggestions are consistent with their views:

1. *Include authentic work.* The work that students include in their portfolio must provide a direct opportunity for them to engage in the types of thinking and abilities typically used by those working in the field or discipline. For example, in a science portfolio students should work on evaluating evidence, using scientific explanations to account for data, or collecting data to support or refute explanations.

2. *Record conceptual development.* Portfolio entries must record students' own explanations, understandings, and conceptual frameworks. This record must be frequently updated as the students progress through a project or a problem solution to show changes in the students' conceptual

framework and thinking as the project develops. It is not enough to include only the finished work. For example, students should periodically record in a science portfolio their current scientific explanation of the events encountered, results observed, and concepts being studied.

3. *Engage in reflective activity.* The students use the portfolio contents as a basis for discussions with the teacher about their understanding of concepts, principles, and theories that underlie the work. The teacher guides the discussion so that students use the same thinking strategies and abilities used by workers in the fields or discipline. For example, if students are working on a scientific problem, they should use the portfolio contents to engage in scientific thinking and activities. The students should record changes in their explanations as new evidence accumulates.

Using these portfolios requires that you have significant knowledge, skill, and ability. Also, you need to be very well versed in the discipline for which you are assessing progress. You should notice, too, that this type of portfolio assessment activity is considerably more spontaneous and less formal than assessing with best works portfolios. These characteristics are not necessarily weaknesses because the portfolio is used with interactive instruction and as a formative evaluation tool.

Best Works Portfolio Organization and Contents

A best works portfolio is organized around learning targets, too. For example, a portfolio may be designed to assess learning targets in the areas of problem solving, mathematical reasoning, mathematical communication, and understanding the core curriculum concepts. Thus each portfolio must contain examples of mathematics investigations, applications, solutions to nonroutine problems, projects, interdisciplinary problem solutions, and writing about mathematics.

Students prepare entries throughout the year. There is no mandated time schedule except the date on which the portfolio is due. However, to help the students prepare, teachers give students explanations and suggestions for deciding which examples the student should include. Figure 12.18 shows some of the explanations and self-reflection questions a teacher gave to students preparing their best works mathematics portfolios.

To be effective, portfolios should emphasize the same standards, curriculum goals, and learning targets emphasized in your daily instruction. The criteria used to evaluate students' portfolio entries should be the same as those used in daily instruction. If your teaching emphasizes students taking responsibility for their own learning, the portfolio procedure you use should be consistent with this approach (Arter & Spandel, 1992)

A portfolio can quickly become a mess of materials and papers that is difficult to assess. To improve the situation, each entry should have an appropriate portfolio entry sheet (or caption) containing the following information:

- Name of the student.
- Date of entry.
- Title or description of the entry. For example, "Comparison of the Population Growth of Canada and the United States."
- Some indication of the learning target or purpose for including the entry. A student may write, "This entry shows that I can use numbers in real-world situations to draw conclusions about how populations grow. I can use growth rates and draw conclusions about when the two populations will be the same."
- Why this particular entry is important or valuable. For example, "I think this was a good piece to include because it shows an actual situation in which I had to use mathematics. Population growth is a social studies topic that I applied mathematics to solve. Also, I had to use a computer spreadsheet program to make the calculations many times in order to discover that the two countries will have the same population in about 59 years."

The size of a portfolio is no small matter! A portfolio that contains too many entries is difficult to understand and may be confusing to students who can get lost in the mass of materials. Also, evaluating long portfolios is difficult and time-consuming.

Portfolio size is related to validity and reliability. Does the portfolio represent the student's attainment? How many entries and what varieties of entries are needed to ensure a representative sample of the student's work? Will a long portfolio be scored less consistently than a short portfolio?

FIGURE 12.18 **Example of suggestions given to fourth-grade students on how to prepare their mathematics portfolios.**

WHAT WORKS?

Here are the types of pieces you should include in your portfolio:

- investigations—studying a mathematical topic or doing a mathematical experiment
- applications—using mathematics to solve real world problems
- non-routine—combining or inventing problem-solving strategies to arrive at solutions or results
- projects—completing problems that take several days or longer
- interdisciplinary—using mathematics with other subjects
- writing—writing about mathematics to explain your thinking or solution

These pieces should also show that you can do these things:

- understand ways to solve problems and do more with the problem (problem solving)
- think by using mathematical ideas and prove your solution is correct by using logical explanations (reasoning)
- explain mathematics to others using mathematical language, symbols, and drawings (mathematical communication)
- understand mathematical topics and use mathematics in other subjects and everyday life (understanding/connecting core concepts)

CHECK IT OUT!

Ask yourself these questions when choosing pieces for your portfolio:

- Did I solve the problem in different ways?
- Have I done other things with the problem?
- Is my answer correct and does it make sense?
- Did I use correct mathematical language, symbols, and/or drawings?
- Is my mathematics connected to other subjects and everyday life?
- Have I listed the mathematical tools (calculators, blocks, beans, etc.) I used?
- Did I explain my thinking and show all my work?
- Does my explanation show that I understand mathematics?
- Have I edited and corrected my work so this is my best effort?
- Have I chosen different types of pieces for my portfolio?
- Did I show all the mathematical topics (core concepts)?
- If I chose a group entry, did I include my own ideas and explanations?
- Have I talked with my teacher about the pieces in my portfolio?

Source: From *Portfolios and You* (p. 3), by the Kentucky Department of Education, 1993, Frankfort: Office of Assessment and Accountability. Reprinted by permission.

Self-Reflection on Portfolio Entries

Effective self-reflection can enhance student learning. The reflection must be substantive (not simply comments like "I worked hard") because that requires students to reason with the subject matter. Reflection also develops metacognitive skills. Arter and Spandel (1992) suggest asking students the following types of questions to prompt self-reflective activities:

- What is the process you went through to complete this assignment? Include where you got ideas, how you explored the subject, what problems you encountered, and what revision strategies you used.
- What were the points made by the group as it reviewed your work? Describe your response to each point—did you agree or disagree? Why? What did you do as the result of their feedback?
- What makes your most effective piece different from your least effective piece?
- How does this activity relate to what you have learned before?
- What are the strengths of your work? What still makes you uneasy?

Although such questions prompt students to review and evaluate their work, the list does not comprise an assessment method per se. Keep in mind that self-reflection is a mental activity. Your assessment of this activity must, therefore, be indirect. Further, although self-reflection appears to be a worthwhile instructional activity, educationally it is not clear that it is either desirable or appropriate to assess formally students' ability to do these self-reflective activities. They may best be handled as informal formative evaluation. In Appendix F we discuss a related area, *metacognition*. Figure F.1 illustrates a student self-assessment questionnaire regarding different aspects of metacognition. You may want to adapt this questionnaire to assess portfolio self-reflection.

Six Steps for Crafting a Portfolio System

Because portfolios are used for such a wide range of formative and summative purposes, a single set of design guidelines is difficult to devise. The six steps that follow are general enough, however, to give you overall guidance in the portfolio-crafting process. Feel free to adapt the steps to suit your particular purposes. The steps express an assessment point of view, namely that assessment should be highly aligned to curriculum and teaching.

Following each step is a set of portfolio-crafting questions to sharpen the focus of your development efforts. Notice that after answering the questions in Step 1, you may decide *not* to develop a portfolio system. Steps 2 through 5 assume that you have completed Step 1 and have decided to use a portfolio system. If you decide to develop a portfolio system, the answers to the questions in Step 1 will set the boundaries and context as you apply the last five steps.

Step 1. Identify Portfolio's Purpose and Focus

- Why do I want a portfolio?
- What learning targets and curriculum goals will it serve?
- Will other methods of assessment serve these learning targets better?
- Should the portfolio focus on best work, growth and learning progress, or both?
- Will the portfolio be used for students' summative evaluation, formative evaluation, or both?
- Who should be involved in defining the purpose, focus, and organization of the portfolio (e.g., students, teachers, parents)?

Step 2. Identify the General Achievement Dimensions to Be Assessed

- Do I need to use the same content and thinking processes framework as I do for individual performance tasks?
- Should I focus primarily on how well the students use the portfolio to reflect on their progress or growth?
- What kinds of knowledge, skills, and abilities will be the major focus of the portfolio?
- If I require a growth portfolio, what do I want to learn about students' self-reflections?

Step 3. Identify Appropriate Organization

- What types of entries (student products and activity records) will provide assessment information about the content and process dimensions identified in Step 2?
- What should the outline or table of contents for each portfolio contain?
- Define each category or type of entry:
  - Which content and process dimension does it assess?
  - What will the teacher or the student "get out of" each entry?
  - What is the time frame for each entry being put into the portfolio?
  - When will the entries be evaluated?
  - What are the minimum and maximum numbers of entries per category?
  - How will the entries within students' portfolios be organized?
  - Will this set of entries fully represent the students' attainment or growth and learning progress?
  - What type of container will I need to hold all of the students' entries, and where will I keep them?

Step 4. Portfolio's Use in Practice

- When will the students work on or use their portfolios (e.g., 15 minutes of every class period)?
- How will the portfolio fit into the classroom routine?
- Will the teacher, student, or both decide what to include in the portfolio?
- Do I need to create a special climate in the classroom to promote the good use of portfolios?
- When will the students and/or the teacher review and evaluate the portfolios?
- How will the portfolios be weighted, if at all, when the time comes to assign letter grades for the marking period?
- Will I schedule a conference to go over the portfolio with the students?
- Will the portfolio be shared with parents? Other teachers? Other students?

Step 5. Evaluation of Portfolios and Entries

- Are scoring rubrics already available for each type of entry?
- Does an evaluation framework or general scoring rubric exist for each type of entry?

- Are the general and specific rubrics aligned with the state standards and school district's curriculum framework?
- Will students, teachers, or both evaluate entries? Which ones?
- Will evaluations of every entry count toward a marking-period grade?
- Given its purpose, is it necessary to have an overall score for the portfolio?
- Should the rubric be holistic, analytic, or annotated holistic?
- Who will score the portfolio (e.g., student, teacher, outsider)?
- How often will the whole portfolio need to be scored (e.g., each week or each marking period)?
- Does an evaluation framework or general scoring rubric exist for evaluating the portfolio as a whole?

Step 6. Evaluation of Rubrics

- Are scoring rubrics available that are consistent with the purpose of the portfolio? With the way each individual entry was evaluated? With the overall curriculum framework?
- Has the scoring rubric been tried on portfolios from different students? From students with different teachers? With what results?
- Does the scoring rubric give the same results for the same students when applied by different teachers?

Electronic Portfolios

Textbook publishers and software developers have created products that allow a portfolio to be presented digitally. These are called **electronic portfolios**. A digitized portfolio can reside on a local computer, a compact disk (CD), or a Website. The software provides an organization for the portfolio contents. Persons then add electronic documents and images in various categories. These digital formats allow for a much wider range of portfolio entries than is possible for portfolios that are housed in folders or crates. In theory, this should enhance validity because more forms of evidence are possible. However, a digital format does not guarantee appropriate learning targets or assignments and scoring schemes that reflect those targets well. Assessment quality principles apply to electronic portfolios, too.

CONCLUSION

This chapter has described a broad range of performance-assessment tasks and scoring schemes. We have presented examples and suggestions for each. This completes our description of how to design and construct or write various assessment methods. Next we turn to how to prepare your students for assessment.

EXERCISES

1. Apply the ideas in Figure 12.2 to a subject you teach or plan to teach. For each category and subcategory, describe one performance assessment applicable to your subject. (Do not use the examples given in the text, but you can adapt them.) You do not have to actually create a workable task. Rather, in one or two sentences describe a task that could be created. Which types of tasks are not applicable to your teaching situation? Explain.
2. Make three columns on a sheet of paper. In the first column list the task properties from Figure 12.7. Select two performance tasks from either your own experience or from this chapter. Identify the second column with one of these two tasks; the third column with the other. Then, in each cell of the table, describe the task with respect to each property.
3. For a subject you teach (or plan to teach), identify learning targets that would be appropriately assessed with on-demand performance tasks using a paper-and-pencil format, and with on-demand performance tasks not using paper and pencil.
   a. Using these results, create one on-demand performance task using paper and pencil and one on-demand performance task not using paper and pencil.
   b. Exchange your tasks with another student in the course. Evaluate each other's task by applying the checklist for judging the quality of performance tasks. If a task resulted in a no answer to one of the checklist items, explain why. Revise the tasks where necessary.
   c. Share your results with others in the course.
4. Select one performance task that you created in Exercise 1, or that you obtained from other sources.

Following the procedures in this chapter, prepare general and specific scoring rubrics. (You may use the framework in Appendix E or another one that your instructor approves.)

a. Write a description of each step you used to craft the rubrics.
b. Exchange your rubrics with another student in the course. Evaluate each other's specific rubrics using the checklist for judging the quality of scoring rubrics. Whenever your rubrics received a "no" answer to one of the checklist items, explain why. Revise the rubrics where necessary.
c. Share your results with others in your course.

5. Select two or more learning targets that can be assessed by one performance task and a corresponding specific scoring rubric (of your own or others' creation). Justify your selection by explaining how this task best assesses these learning targets.
   a. Administer the performance task to at least five students. Score the task using the scoring rubric.
   b. Write a short essay describing your scoring experience. Was the scoring rubric adequate? Were there any reliability problems in using it? Why or why not? Make suggestions for improving the scoring rubric based on your experience.
   c. Prepare a summary of your students' results.
6. Design a best works portfolio system for assessing students in the subject you teach (or plan to teach). Follow the six-step procedure suggested in the chapter.
   a. Prepare a report describing the portfolio system you designed. Be sure your report addresses all of the questions listed under each step.
   b. Discuss your portfolio system in class with other students.

Preparing Your Students to Be Assessed and Using Students' Results to Improve Your Assessments

KEY CONCEPTS

1. To prepare students for an upcoming assessment, give students the information and skills they need to perform their best.
2. Testwiseness is the ability to use test-taking strategies, clues from poorly written items, and experience to improve a score beyond that expected from mastery of the subject matter.
3. Test anxiety is increased emotional tension based on a student's appraisal of a testing situation.
4. The assembly and administration of an assessment affect the validity of the scores.
5. Correction for guessing formulas adjust scores for the expected effects of random choices. They are not recommended for classroom use.
6. Item analysis results can be used to improve the quality of true-false, matching, and multiple-choice items. Analogous statistics can be examined for constructed response (multipoint) tasks.
7. Item difficulty shows students' average level of performance on a test item.
8. Item discrimination shows how students' performance on an item is related to their total test performance.
9. Improve multiple-choice item quality by editing items flagged by unacceptable difficulty or discrimination indices or by poorly functioning distractors.
10. Use item analysis information to select multiple-choice items appropriate to the test's purpose.
11. Computers can be a great aid in testing. Storing item analysis information in computerized item banks makes it easier to use. Computer applications can also help make tests accessible to students with disabilities.

IMPORTANT TERMS

ambiguous alternatives

complete versus partial ordering of students

content analysis of the responses

correction for guessing formulas

dichotomous item scoring

homogeneous versus heterogeneous test

item analysis

item bank

item difficulty index (*p* and *p**)

item discrimination index (*D* and *D**)

maximum performance assessment

miskeyed items

negatively discriminating item

nondiscriminating item

poorly functioning distractor

positively discriminating item

relative versus absolute achievement

From Chapter 13 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

task-directed versus task-irrelevant thoughts
test anxiety
testwiseness
types of test-anxious students
typical performance assessment
upper-, middle-, and lower-scoring groups

PREPARING STUDENTS FOR ASSESSMENT

Assess Maximum, Not Typical, Performance in the Classroom

You should assess students' maximum performance rather than their typical performance (Cronbach, 1990). You assess **maximum performance** when you set the conditions so that students are able to earn the best score they can. You assess **typical performance** when you gather information about what a student would do under ordinary or typical conditions.

For example, you may have taught students a practical skill such as balancing a checkbook, and your assessment procedure gathers information about whether each student is capable of doing so. This is maximum performance assessment. Some students, on the other hand, may make errors later outside class when actually using checks, or they may never reconcile their checking account. Thus, such students may be *capable* of performing the skill you taught, but may *typically not perform* the skill to their maximum capacity. Because schooling usually attempts to teach learners new abilities at high levels, achievement assessments are carried out under conditions that encourage students to perform to the best of their abilities.

Give Students Enough Information Before Assessing Them

We have described informing students about an upcoming assessment and about how it will be scored as a professional responsibility. To assess students under the best conditions, you should provide at least the following information about your upcoming assessment:

1. When it will be given.
2. The conditions under which it will be given (timed, speeded, take-home test).
3. The content areas it will cover.
4. The emphasis or weighting (point value) of content areas to be included on the assessment.
5. The types of performance the student will have to demonstrate (the kinds of items on the test, the degree to which memory will be required).
6. The way the assessment will be scored and graded (e.g., will partial credit be given?).
7. The importance of the particular assessment result in relation to decisions about the student (e.g., it will count for 20% of the marking period grade).

When an Assessment Will Be Given If you want students to perform at their best, you need to tell them when your test will be given so they can prepare in advance. Students, particularly those taking courses taught by more than one teacher, need to organize their study efforts and set their priorities. They can learn to do this planning when they know the test date in advance. Teachers of various subjects should coordinate their schedules of assessment so they are spread out. However, the end of the marking period is often problematic.

Pop Quizzes Do Not Assess Maximum Performance Some teachers advocate "surprise" or "pop" quizzes. Their reasoning is often some vague notion that a good student should always be prepared to perform on command. This seems to be an unrealistic expectation of students. Some teachers use surprise quizzes to threaten or to punish a disobedient class. The authors consider this an unethical use of an assessment.

Students with special problems often benefit from knowing about an assessment well in advance. Test anxiety and fear are likely to diminish when a student can rationally plan a program of study for a forthcoming assessment (Mealey & Host, 1992). Children with disabilities mainstreamed in a regular class often have supplemental instruction from an itinerant teacher or tutor who sees them only once or twice a week. Suppose a youngster with a hearing disability has not understood Wednesday's lesson, and the itinerant teacher regularly comes on Monday. Further, suppose the quiz is "popped" on Friday. How can this youngster be expected to plan effectively and use the resources provided when the regular teacher is unpredictable?

Assessment Conditions Tell students the conditions under which they are expected to perform: How many items will be on the test? How much time will the students have to complete the assessment? Will the assessment be speeded? Will it be open or closed book? Will there be a penalty for guessing? And at what time of day will it be given (if not during a regular period)?

Explain to Students What the Test Will Include Saying that the assessment will cover the first three chapters of the book doesn't help students much. To plan and study effectively, students need more detail. Some teachers prepare lists of study questions to help students focus their efforts. This may be especially helpful for elementary students for whom almost everything in a book seems to be equally important. Study questions also help older students, especially when a large amount of material has been covered during the term. For high school and college students, an alternative to developing a set of study questions is to give them a copy of the assessment blueprint (see Chapter 6), a list of learning targets, a copy of the scoring criteria (or rubrics), or a detailed content outline indicating the number of items covering each element.

Explain What the Test Will Emphasize Tell students how the content in an assessment is weighted, including how many items (and how many points) will be devoted to each objective, content element, or blueprint cell. Weight of the different parts of an assessment should match your teaching emphasis; otherwise, the results will have low validity. Students can waste hours studying a topic that will be of little or no importance on the assessment. Many teachers share their assessment plan with students, telling them at the beginning of the course or marking period the weight they assign to each assignment, quiz, test, and classroom performance activity. Students can then organize their efforts in terms of these priorities.

Give Opportunity to Practice Expected Performance Give students the opportunity to practice the kind of performance for which you will hold them accountable. Unfortunately, students frequently have to guess at the nature or type of question that will appear on an assessment. For example, a teacher gave a sixth grader practice exercises that asked him to identify prepositional phrases in isolation using a given list of words and phrases. The next day, his assessment consisted of finding the subject, predicate, and prepositional phrase in the more authentic context of several paragraphs. Unfortunately, he never had the opportunity to practice the task for which he was held accountable.

The best way to familiarize students with tasks that will appear on an assessment is to give them sample tasks, perhaps an old form of an assessment on which they can practice. This may be particularly effective when the types of tasks to appear on the assessment are complex and/or unfamiliar to the students.

Tell Students How You Will Score the Test Telling students how you will score the assessment helps them prepare, especially for answering open-ended tasks. If you will assign points for spelling important terms and proper names, then the students need to practice these spellings in addition to learning the main ideas and rehearsing how to organize their answers. Students also need to know whether and how you will award marks for less-than-perfect answers and how much weight (i.e., marks) you will give for each question. Be sure to share scoring rubrics with students well in advance of giving a test.

Tell Students How the Test Results Will Be Used Tell students the importance of the assessment score for any decisions you will make about them, including putting students into groups, placing them in another section of the course, assigning them to remedial instruction, giving them enrichment or advanced work, and assigning grades.

Minimum Assessment-Taking Skills

Skills You Need to Teach Students Students need more than information about what an assessment is: They need to learn how to take tests. You may need to teach students the following minimum assessment-taking skills, perhaps through direct instruction in the classroom (Ebel & Frisbie, 1991):

- Paying attention to oral and written directions and finding out the consequences of failing to follow them.
- Asking how the assessment will be scored, how the individual tasks will be weighted into the total, and how many points will be deducted for wrong answers, misspellings, or poor grammar.

- Writing their responses or marking answers neatly to avoid lowered scores because of poor penmanship or mismarked answers.
- Studying throughout the course and in paced reviewing to reduce cramming and fatigue.
- Using assessment time wisely so that all tasks are completed within the given time.
- Using their partial knowledge and guessing appropriately.
- Reflecting, outlining, and organizing answers to essays before writing; using an appropriate amount of time for each essay.
- Checking the marks they make on the separate answer sheets to avoid mismatching or losing one's place when an item is omitted.
- Reviewing their answers to the tasks and changing answers if they can make a better response.

Avoid Shortchanging Your Students Some teachers have strong opinions about not giving multiple-choice items to students. Others give only short quizzes and tests lasting 15 to 20 minutes. Still others give almost no tests, relying on assignments and take-home work. We ask you to consider your own position on these matters. Students will almost always be required to take state assessments and/or standardized tests. Doing well on these tests will be important for your students because decisions about them and your school will depend on how well they do.

We are not advocates of using multiple-choice tests exclusively, nor even extensively. Neither do we advocate always giving long tests. But we must be fair to the students. If we are expecting them to do well on the state assessments and standardized tests, then they should experience these types of assessment during their normal classes as part of their normal instruction and assessment process. Prepping students for taking these longer multiple-choice tests a week or so before the tests does not seem right. It is a waste of instructional time and may well be an unethical teaching practice.

TESTWISENESS

A Testwiseness Quiz

Before reading further, take the following short test (adapted from Diamond & Evans, 1972, p. 147). Be sure to mark an answer for every item, even if you are unsure of the answer. There *is* a correct or best answer for every item.

1. The Augustine National Party has its headquarters in
   a. Camden, New Jersey.
   b. St. Augustine, Florida.
   c. Palo Alto, California.
   d. Dallas, Texas.
2. Hermann Klavermann is best known for
   a. developing all musical scales used in the western world.
   b. composing every sonata during the Romantic Era.
   c. translating all Russian classics into English.
   d. inventing the safety pin.
3. The Davis Act of the 20th century
   a. provided more money for schools.
   b. struck down an earlier law.
   c. prohibited the manufacture, sale, transportation, or use of several specific drugs that were being used for illegal purposes.
   d. gave a raise to government employees.
4. Harold Stone's book *The Last Friendship* is an example of an
   a. political satire.
   b. autobiography.
   c. science fiction.
   d. biography.
5. The population of Franktown is more than
   a. 50 thousand.
   b. 60 thousand.
   c. 70 thousand.
   d. 80 thousand.

Each item's content is fictitious, but the right answer to each can be determined by using certain clues in the item:

Item 1. An obvious association between a word or phrase in the stem (*Augustine National Party*) and one in an alternative (*St. Augustine*).

Item 2. Specific determiners in the alternatives (*all, every*) result in these being eliminated from consideration.

Item 3. A longer, more qualified answer is keyed as the correct response.

Item 4. A grammatical clue (*an*) is contained in the stem.

Item 5. An alternative overlaps or includes the others.

A Taxonomy of Testwiseness Skills

The ability to correctly answer items like the preceding is often called testwiseness. **Testwiseness** is the ability to use assessment-taking strategies, clues from poorly written items, and experience in taking assessments to improve your score beyond what you would otherwise attain from mastery of the subject matter itself. When you write classroom assessments, be aware of how students may take advantage of your idiosyncrasies in item writing or flawed items to improve their scores without attaining the desired level of mastery. Figure 13.1 is an outline or taxonomy of testwiseness principles.

You should create good-quality assessments that minimize any advantage that testwise students have. It will be beneficial, however, if you teach all students many of the skills listed in Part I

FIGURE 13.1 **A taxonomy of testwiseness principles.**

I. Elements independent of test conductor or test purpose.

A. Time-using strategy.

1. Begin to work as rapidly as possible with reasonable assurance of accuracy.
2. Set up a schedule for progress through the test.
3. Omit or guess at items (see I.C. and II.B.)
4. Mark omitted items, or items which could use further consideration, to assure easy relocation.
5. Use time remaining after completion of the test to reconsider answers.

B. Error-avoidance strategy.

1. Pay careful attention to directions, determining clearly the nature of the task and the intended basis for response.
2. Pay careful attention to the items, determining clearly the nature of the question.
3. Ask examiner for clarification when necessary, if it is permitted.
4. Check all answers.

C. Guessing strategy.

1. Always guess if right answers only are scored.
2. Always guess if the correction for guessing is less severe than a “correction for guessing” formula that gives an expected score of zero for random responding.
3. Always guess even if the usual correction or a more severe penalty for guessing is employed whenever elimination of options provides sufficient chance of profiting.

D. Deductive reasoning strategy.

1. Eliminate options which are known to be incorrect and choose from among the remaining options.
2. Choose neither or both of two options which imply the correctness of each other.
3. Choose neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other.
4. Restrict choice to those options which encompass all of two or more given statements known to be correct.
5. Utilize relevant content information in other test items and options.

II. Elements dependent upon the test constructor or purpose.

A. Intent consideration strategy.

1. Interpret and answer questions in view of previous idiosyncratic emphases of the test constructor or in view of the test purpose.
2. Answer items as the test constructor intended.
3. Adopt the level of sophistication that is expected.
4. Consider the relevance of specific detail.

B. Cue-using strategy.

1. Recognize and make use of any consistent idiosyncrasies of the test constructor which distinguish the correct answer from incorrect options.
   a. He makes it longer (shorter) than the incorrect options.
   b. He qualifies it more carefully, or makes it represent a higher degree of generalization.
   c. He includes more false (true) statements.
   d. He places it in certain physical positions among the options (such as in the middle).
   e. He places it in a certain logical position among an ordered set of options (such as the middle of the sequence).
   f. He includes (does not include) it among similar statements, or makes (does not make) it one of a pair of diametrically opposite statements.
   g. He composes (does not compose) it of familiar or stereotyped phraseology.
   h. He does not make it grammatically inconsistent with the stem.
2. Consider the relevancy of specific detail when answering a given item.
3. Recognize and make use of specific determiners.
4. Recognize and make use of resemblances between the options and an aspect of the stem.
5. Consider the subject matter and difficulty of neighboring items when interpreting and answering a given item.

Source: From “An Analysis of Test-Wiseness,” by J. Millman, C. H. Bishop, and R. L. Ebel, 1965, *Educational and Psychological Measurement,* 25, pp. 711–713. Copyright © 1965 by Sage Publications. Reprinted by permission of Sage Publications. Inc.

of Figure 13.1 so they are not at a disadvantage when being assessed with more testwise peers. And of course, you should work to make sure your own tests are well crafted, so the "skills" in Part II do not help with answers. Research has demonstrated that testwiseness is learned, and it improves with grade level, experience in being assessed, maturation, and motivation to do well on the assessment (Geiger, 1997; Sarnacki, 1979).

Advice About Changing Answers

Will students benefit if they change their answers once they have been marked on the answer sheet? Despite popular opinion, it *does* pay to change answers if changing them is based on a thoughtful reconsideration of the item. A summary of the research findings (Wise, 1996) on this issue follows.

- Most test takers and many educators believe it does not pay to change answers.
- Most students, however, do in fact change their answers to about 4% of the items.
- Research studies show that it does, in fact, pay to change answers. Typically two out of three answers changed will become correct.
- The payoff for changing answers diminishes as the items become more difficult for the student.
- Lower-scoring students benefit less from changing answers than higher-scoring students do.

TEST ANXIETY

Nature of Test Anxiety

Task-Directed and Task-Irrelevant Thoughts

How students perceive being evaluated varies widely and those perceptions affect students' performance on assessments. Some students are motivated to perform well; others don't care. Among the students who are motivated to do well, assessments and evaluations are likely to lead to increased emotional tension: **test anxiety**. Students' perceptions of evaluation situations shape their reactions to them. Some well-motivated students may perceive these evaluation situations as challenges, whereas other equally well-motivated students perceive them as threats. A student who perceives an assessment as threatening may not have the ability to perform the task at hand, not have been taught how to perform the task, or not have properly studied or otherwise prepared for the assessment (Benson, 1989). Not all perceived threats are based upon poor preparation, however.

Students who accept assessments and evaluations as challenges have thoughts that are **task-directed**. Their thoughts and actions are focused on completing the tasks and thereby reduce any tensions that are associated with them. Schutz, Distefano, Benson, and Davis (2004) called these task-focusing processes. Students who perceive assessments and evaluations as threats have **task-irrelevant thoughts**: They are self-preoccupied, centering on what could happen if they fail, on their own helplessness, and sometimes on a desire to escape from the situation as quickly as possible. Schutz et al. (2004) called these emotion-focusing processes.

Cognitive appraisal—that is, students making judgments about the tests they take and about their ability to manage the situation—affects how students cope with text anxiety (Schutz, et al., 2004). Emotional reactions to an assessment situation trigger worry, which in turn results in poor performance. That is, highly test-anxious students worry about doing poorly. This keeps them from focusing their attention on the task at hand. If students can change their appraisal of the situation, however, they can also change their emotional experience and focus.

Factors in Test Anxiety Sarason (1984) conceptualized students' reactions to assessment situations as four related factors: tension, worry, test-irrelevant thinking, and bodily reactions. Tension is the feeling of unease or jitters before a test. Worry includes worrying about failure and what is going to happen. Test-irrelevant thinking, as discussed above, is thinking about things other than the test, which in turn interferes with performance. Bodily reactions include headaches, upset stomach, and rapid heartbeat.

Davis and Li (2008) suggest that it might be helpful to consider students' emotional reactions to tests more broadly than just anxiety reactions. Examples of students' potential beliefs or judgments about tests are listed below, according to what emotion they engender: anxiety, anger, or pride. Beliefs and judgments that lead to pride and self-confidence are more productive than beliefs that lead to anxiety or anger.

Example

Students' beliefs and judgments about tests

Anxiety	*Anger*	*Pride*
Tests are important	Tests are important	Tests are important
Tests are ***not*** helping my goals	Tests are ***not*** helping my goals	Tests are helping my goals
Tests scores are ***not*** under my control *(my fault)*	Tests scores are ***not*** under my control *(someone else's fault)*	Tests scores are under my control
I ***cannot*** cope with problems on tests	I ***cannot*** cope with problems on tests	I can cope with problems on tests

Source: Davis, H. A., & Li, J. (2008). *The relationship between high school students' cognitive appraisals of high stakes tests and their emotion regulation and achievement*. Paper presented at the annual meeting of the American Edcational Research Association, New York. Used by permission.

Three Types of Test-Anxious Students

There are at least three **types of test-anxious students** (Mealey & Host, 1992). Your ability to recognize these differences among students will help you work with them so they perform their best on your assessments. First are students who do not have good study skills and do not understand how the main ideas of the subject you are teaching are related and organized. These students become anxious about an upcoming evaluation because they have not learned well (Culler & Holahan, 1980; Naveh-Benjamin, McKeachie, & Lin, 1987). The second group contains students who do have a good grasp of the material and good study skills but have fears of failure associated with assessment and evaluation (Herman, 1990; McKeachie, Pollie, & Spiesman, 1985). Third are students who believe they have good study habits but who do not. They perform poorly on assessments and learn to be anxious about being assessed (Mealey & Host, 1992).

Helping Test-Anxious Students

The following eight factors were shown to be related to test anxiety (Hembree, 1988) and may be under your control in classroom assessment situations:

1. When students perceive an assessment to be difficult, their test anxiety rises.
2. At-risk students have higher levels of test anxiety than passing students.
3. Students whose teachers gave them item-by-item feedback after the test have lower test anxiety than students who receive no feedback.
4. Tests whose items were arranged from easy to difficult raise test anxiety less than tests with other item arrangements (Tippets & Benson, 1989).
5. More frequent testing of highly test-anxious students seems to improve their performance.
6. Highly test-anxious students are more easily distracted by auditory and visual activity than less test-anxious students.
7. Giving extremely test-anxious students instructions to concentrate their attention on the assessment tasks and not to let themselves be distracted from the tasks is more beneficial to their performance than simply reassuring them with "don't worry" or "you'll be fine" statements (Sarason, 1984).
8. Students with low test-taking skills can lower their test anxiety with testwiseness training.

In addition, Mealey and Host (1992) suggest that you ask your students what you might do to help them feel more relaxed or less nervous before, during, and after you assess them. The researchers' own developmental reading college students reported these four suggestions:

1. The teacher should not talk or interrupt while students are working on an assessment.
2. The teacher should review the material with the entire class before the assessment is given.
3. The teacher should not walk around looking over students' shoulders while they are being assessed.
4. The teacher should convey a sense of confidence about students' performance on an upcoming assessment (and avoid such statements as "This is going to be a difficult test").

Further reviews of test anxiety and its treatment can be found in Ergene (2003), Hembree (1988), and Zeidner (1998).

ASSESSMENT FORMAT AND APPEARANCE

The final appearance and arrangement of your test are important to the validity of the results. An illegible, poorly typed, or illogically arranged assessment annoys the well-prepared student, can cause unnecessary errors, and gives all students the impression that you have not taken your assessment responsibilities seriously. The organization and appearance of an assessment may be especially important for less able students.

As a rule, you should type a test and duplicate it so that each student can have a copy. (Obvious exceptions are dictated spelling assessments and similar assessments of aural abilities.) Sometimes a teacher will write the items on the board or dictate them to the class. If you do this, it may cause problems for students, especially those with visual, listening comprehension, or hearing problems. If you dictate the questions, you use valuable time that your students could otherwise spend in responding to the items. Further, reading a question aloud and requiring students to write their responses places a demand on short-term memory that many students cannot meet.

Test Layout and Design

Experts usually recommend placing items in the order of difficulty, with the easiest items first. If items are grouped by type of format, arrange them from easiest to most difficult within each format. Most students can go through the easiest items quickly and reserve the remaining test time for the difficult items. Some research shows that the easiest-to-hardest item arrangement reduces anxiety and increases performance (Tippets & Benson, 1989). Another way to arrange items is according to the sequence in which the content was taught or appeared in the textbook. Students can then use this subject-matter organization as a kind of "cognitive map" through which they can retrieve stored information. If you use this sequential arrangement, you should tell students to skip over difficult items and go on to subsequent items, which may be easier. Always encourage students to return to the omitted items if they have time. Better yet, within content areas arrange the items from easiest to most difficult. This minimizes test-created anxiety and, in turn, raises the validity of your assessment results.

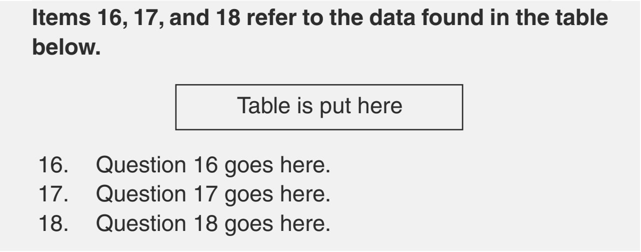
Directions to Students

Assessment directions should contain certain minimum information: the number and format of items, amount of time allowed for the assessment, where and how answers should be written, any correction or penalty for guessing, and the general strategy the student should follow when answering questions. For example, should students guess if they think they know the answer but are unsure? Should they answer all items or should they omit some? Which ones? Should they do each item in turn, or should they skip those they are uncertain about, returning to them later if they have time? If the student perceives that the answer to an item requires an opinion, whose opinion is being asked? Most written directions need not be elaborate.

Use side headings on the pages with the test questions to signal a change in the general directions that may occur within the test booklet. Some items may require specific, rather than general, directions. Here is an example:

Example

Example of side heading on a test page to signal changes in the general directions



Items 16, 17, and 18 refer to the data found in the table below.

Table is put here

16. Question 16 goes here.
17. Question 17 goes here.
18. Question 18 goes here.

Make sure your test copies are clear and readable. Poor-quality copies may affect student performance. Test security may be a problem if the tests are sent to a central location for duplication and assembly. Be sure to check the security procedures in your school.

Preparations for Scoring the Assessment

Prepare for scoring an assessment before you administer it. Prepare and *verify* every answer on the scoring key in advance, so that you can score students' assessments efficiently and accurately, and report results to the students quickly for feedback and motivation. Advance preparation of the answer key will help you identify errors in the assessment items, too.

Separate answer sheets are not recommended for the first three grades. However, with older elementary and high school students, it may be advisable to use a separate answer sheet for objective items. This greatly facilitates scoring and permits the test booklet to be reused. It also gives students practice for the state assessment or other standardized tests they will have to take. An answer sheet for completion items might consist of columns with numbered blanks, each number corresponding to an item number. The student writes the answer to an item on the correspondingly numbered blank. If you have essays or extended responses, be sure to provide an answer sheet or examination booklet to record students' responses.

Scanning the Answer Sheets

Scannable answer sheets are available in most schools. These can be used for hand scoring as well as scanning. If you are planning to scan, check with the office in your school that does the scanning to be sure that you use the proper answer sheets. Also be sure the students follow the correct answer-marking procedures. You can make your own scoring key by punching out the correct answers on an answer sheet. Lay the punched sheet on top of each student's answer sheet to score it.

CORRECTION FOR GUESSING

Correction for Guessing Formulas

With true-false and/or multiple-choice items, **correction for guessing formulas** are sometimes applied to scores by subtracting from the number of right answers a fraction of the number of wrong answers. An astute student may wonder how the machine can "get into my head" and figure out whether the student guessed. Assure your student no machine can do that. What the correction for guessing formulas do is correct scores so that *on average* the effects of chance (and therefore the probability of getting a correct answer by guessing) are removed. Here is the usual formula:

$$\text{corrected score} = R - \frac{W}{(n-1)} \qquad \text{[Eq. 13.1]}$$

where

R means the number of items answered correctly
W means the number of items marked wrongly
n means the number of options in each item

If there are two choices per item (e.g., true-false), then

$$\text{corrected score} = R - W$$

If there are four options per item, then

$$\text{corrected score} = R - \frac{W}{3}$$

The correction formula is designed to eliminate the advantage a student might have as a result of guessing correctly. Here is an example of how to use the correction. You would apply the correction to every student.

Example

How to apply the correction for guessing correctly score formula

Suppose Juan took a 50-question multiple-choice test with four options per item. Further, suppose Juan's test results were 40 items marked correctly, 6 items marked wrongly, and 4 items omitted. Applying the formula, we find:

$$\text{corrected score} = \mathrm{R} - \frac{\mathrm{W}}{(\mathrm{n}-1)}$$

$$\text{corrected score} = 40 - \frac{6}{(4-1)} = 40 - 2 = 38$$

Notice the number of omitted items is not used in this correction formula; only the number of answers marked wrongly (W) and the number marked correctly (R).

A complementary version of the preceding correction formula does use the number of omitted items: Instead of penalizing a student for responding wrongly, it rewards the student for omitting items (i.e., for refraining from guessing). This formula is

$$\text{adjusted score} = R + \frac{O}{n} \qquad \text{[Eq. 13.2]}$$

where

R means the number of items answered correctly
O means the number of items omitted
n means the number of options in each item

(The term *adjusted* instead of *corrected* distinguishes this formula from the previous one. This is not standard practice. The general term for such equations is formula scoring.) Here is an example of its use:

Example

How to apply the correction for guessing adjusted score formula

Suppose Juan took a 50-question multiple-choice test with four options per item. Further, suppose

Juan's test results were 40 items marked correct, 6 items marked wrong, and 4 items omitted. Applying the formula, we find:

$$\text{adjusted score} = R + \frac{O}{n}$$

$$\text{adjusted score} = 40 + \frac{4}{4} = 41$$

This formula credits the student with the number of points to be expected if random responses were substituted for the omitted responses. If a student omitted every item, the score would be equal to the average score expected if the student guessed randomly on every item. Thus, the scores obtained by the adjusted score formula will be higher than the same students' scores if they had been obtained from the corrected score formula. However, the scores under the two methods are perfectly correlated; that is, the rank ordering of persons is the same regardless of which formula is used.

The uncorrected score (R) is simply the number of items marked correctly. When every student marks every item, the uncorrected scores are perfectly correlated with the corrected or adjusted scores so that the rank ordering of persons is the same, whether or not the scores are corrected for guessing.

Figure 13.2 lists a few things to keep in mind when deciding whether to use a correction formula. On balance, we recommend that you do not use formula scoring for most classroom assessment purposes.

Current Practices Among Test Publishers

Test publishers disagree about the use of correction for guessing. Most current commercial achievement tests use item response theory, whose mathematical models take guessing into account in scoring, without using Equations 13.1 and 13.2.

If you hand-score a standardized test, follow the instructions in the manual exactly. If hand scoring is required and you fail to apply a correction the publisher intended, apply it when the test publisher didn't intend it to be used, or otherwise alter the instructions to students at the time of testing, you will make the test's norms unusable because the alterations result in new unstandardized test conditions.

ITEM ANALYSIS FOR CLASSROOM ASSESSMENTS

Item analysis is the process of collecting, summarizing, and using information from students' responses to make decisions about each item. Standardized test developers, especially developers of norm-referenced tests, try out as many as five times more items than will appear on the final version of a test. Item analysis data from these tryouts are used to help select items for the final form. The developers discard items that fail to display proper statistical properties. Your classroom assessments, being more closely linked to the daily teaching-learning process, serve purposes that are somewhat different from published standardized tests. Thus, you will use item analysis data differently than a test publisher.

FIGURE 13.2 Things to consider for correction for guessing.

1. A correction formula does not correct for good luck nor compensate for bad luck.
2. The relative ordering of pupils is usually the same for uncorrected as for corrected scores.
3. The chance of getting a good score by random guessing is very slim.
4. Pupils who want to do well on the test, and who are given enough time to attempt all items, will guess on only a few items.
5. Encouraging pupils to make the best choice they can, even if they are not completely confident in their choice, does not seem to be morally or educationally wrong.
6. Responding to an item on a rational basis, even when lacking complete certainty of the correctness of the answer, provides useful information on general educational achievement.
7. Using a correction-for-guessing penalty may discourage slower students from guessing blindly on items near the end of a test when time is short.
8. Correction-for-guessing directions do not seem to discourage the test-wise or risk-taking examinee from guessing, but do seem to discourage the reluctant, risk-avoiding, or non-test-wise examinee.
9. A formula score makes the scoring more complicated, offering additional opportunities for the teacher to commit scoring errors.

Source: From *Measuring Educational Achievement* (pp. 251–257), by R. L. Ebel, 1965, Englewood Cliffs, NJ: Prentice Hall. Reprinted by permission.

Classroom Uses for Item Analyses

For teacher-made assessments, the following are among the important uses of item analyses:

1. *Determining whether an item functions as you intended.* You can't expect to write perfectly functioning items. To decide whether an item for a classroom assessment is functioning properly, you need to know whether it assesses the intended learning targets, whether it is at the appropriate level of difficulty, whether it distinguishes those who have command of the learning targets from those who do not, whether the keyed answer is correct, and (for response-choice items) whether the distractors are working. Procedures to help you decide whether an item seems to be assessing the intended learning target were discussed in Chapter 6. The other four elements are discussed in this chapter.

2. *Feedback to students about their performance and as a basis for class discussion.* Students are entitled to know how their performance on each assessment task is marked and the correct answer to each task. Going over a test with students makes instructional common sense: You can correct students' errors, you can clarify for students the level of detail you expect of them, and you can reinforce good (and correct) responses. Also, students lacking testwiseness skills may learn how a correct answer is formulated or why (in response-choice items) foils are incorrect, and you can alleviate some test anxiety if you teach your students to view your assessments rationally in the context of instruction.

3. *Feedback to the teacher about pupil difficulties.* A simple procedure such as tabulating the percentage of students answering an item correctly may provide you with information about areas that need additional instruction and remediation. Many school systems have electronic equipment to scan answer sheets and a computer program that can provide an item analysis. Feedback from item analysis can help you focus your teaching on both group and individual needs. Note, however, that a subscore based on a cluster of several items measuring similar learning targets provides more reliable information than does a single item, so use these results cautiously.

You will also find it helpful to identify the nature of students' errors on assessment tasks. With essay, short-answer, and completion items, a **content analysis of the responses** will determine the major types of student errors and how often they occur.

4. *Areas for curriculum improvement.* If particular content is repeatedly difficult for students, or if certain kinds of errors occur often, perhaps the problem extends beyond you: A more extensive curriculum revision may be needed. Item analysis data help to identify specific problems. But any assessment is likely to represent a school's curriculum objectives incompletely, so you should use caution when attempting to generalize item analysis to the whole of student learning.

5. *Revising the assessment tasks.* Use information about students' responses to and perceptions of an item to revise it. Items can be reused for future assessments and, if you revise a few each time, the overall quality of the assessment will eventually improve. Usually you will find it less time-consuming to revise an item than to write a new one. Some teachers, especially in the junior and senior high schools (and in colleges), develop an item file or **item bank**. They write and try new items, and through item analysis they keep the best items each time, revise some, and discard the rest. You can keep a copy of the item and information about it on a card in a file for future use. Following is an example of one card in such a file:

Example

Example of an item card with item analysis data for one item. Tabulations were made for Item 1 on the summary record form shown in Figure 13.3.

Course: English 10

Date(s): Fall 2007
Spring 2008
Fall 2008

Topic: Poetry

ITEM

The poet John Donne began the second verse of his poem "The Message" with this line: "Send home my harmless heart again." This is an example of what element of poetry?

*a. Alliteration
b. Assonance
c. Irony
d. Simile

ITEM DATA SUMMARY

	Upper Group	Lo wer Group	Middle Group
*a.	10	7	14
b.	0	0	0
c.	0	2	0
d.	0	1	0
Omits	0	0	0

Difficulty Index: 0.91 **Number of Students:** 34
Discrimination Index: 0.3

Or, you can use a database or spreadsheet program to create your item bank.

After several years, a file of good items accumulates. Once a file of items is established, equivalent versions of a test can be constructed relatively easily. You can construct equivalent versions of tests and use them for makeup tests when persons are absent during the regularly scheduled administration, when you teach multiple sections of the same course, or when you use tests in an alternating pattern from year to year.

6. *Improving item-writing skills.* Probably the most effective way to improve your item-writing skills is to analyze the items and understand the ways in which students respond to them, and then use this information to revise items and try them again with students.

Item Analysis of Response-Choice Tests

The basic bits of data you need to begin an analysis of response-choice items (true-false, matching, or multiple-choice) are the responses each student makes to each item. Although this information is easier to use if students have marked their answers on separate answer sheets, such sheets are not necessary. Here is a summary of the steps necessary for doing an item analysis.

Step 1. Score each student's test by marking the correct answers and putting the total number correct on the test (or answer sheet).

Step 2. Sort all the papers in numerical order according to the total score.

Step 3. Determine the upper, middle, and lower groups.

Step 4. Tabulate the number of students choosing each alternative in the upper and lower groups, and tabulate the number of students in the middle group who chose the correct answer.

Step 5. Calculate the difficulty index for each item.

Step 6. Calculate the discrimination index for each item.

Step 7. Using the results of Step 4, check each item to identify poor distractors, ambiguous alternatives, miskeying, and indications of random guessing.

Although this section is written primarily for analyses of response-choice items, you can use several of the techniques described with any assessment tasks that are **dichotomously scored** (correct/ incorrect or pass/fail), such as completion or short-answer items. Item analysis techniques do exist for analyzing tasks scored more continuously, such as essays or performance assessments, but we start with the simplest case, response-choice tests with right/wrong scoring.

Upper- and Lower-Scoring Groups (Step 3)

After you have scored the tests, arrange them in numerical order according to the students' total score. Next, divide the stack of tests into three groups: **upper-, middle-, and lower-scoring groups**. You then contrast the responses of the upper- and lower-scoring groups in various ways (described later) to determine whether each item is functioning well.

How you form these groups is important. When the total number of students taking your test is between 20 and 40, select the 10 highest-scoring and the 10 lowest-scoring papers (Whitney & Sabers, 1970), but keep the middle-scoring group intact. (When there are 20 students, there will be no middle group.) If there are 20 students or fewer, the responses of only one or two students may greatly influence the results you will obtain from the procedure described here. If you use item analysis with too few students, you may come to quite incorrect conclusions about how a particular item would function if you were to use it again. Nevertheless, if you want to go ahead with the analysis for groups with very few students, separate the test papers into two sets (upper and lower halves) and interpret the results cautiously. For groups larger than 40, testing experts frequently recommend using the upper- and lower-scoring 27% of the group on technical grounds (Kelly, 1939). For purposes of classroom assessment, however, when the group is almost always smaller than 40, any percentage between 25 and 33 seems appropriate.

Summarize Responses to Each Item (Step 4)

For each item, record the number of students in the (a) *upper group* choosing each alternative (and, separately, the number not responding [omitting]); (b) *lower group* choosing each alternative (and, separately, the number omitting the item); and (c) *middle-scoring group* choosing the correct alternative. The item file card example we gave previously shows the results of such tabulation for one item. You also can make a form to record the necessary numbers for several items on a single page, or you can simply write this information in the margin of the teacher's copy of the test.

Without a doubt, the most tedious part of an item analysis is tabulating the students' responses to items. Using an upper and lower group instead of the entire class makes the task easier. One simplifying procedure is to make a form such as in Figure 13.3. Or, you may find that, like many teachers, you would not do an item analysis by hand, but working through an example by hand helps you understand and interpret item analysis printouts.

Compute the Item Difficulty Index (*p*) (Step 5)

The fraction of the total group answering the item correctly is called the **item difficulty index (*p*)**. To compute it, add together the number of students choosing the correct answer in the upper, middle, *and* lower groups, then divide this sum by the total number of students who took the test. Equation 13.3 summarizes this.

$$p = \left[\frac{\text{number of students choosing the correct answer}}{\text{number of students taking the test}}\right]$$

$$= \left[\frac{\begin{array}{c}\text{number of students choosing the correct answer}\\ \text{for the upper} + \text{middle} + \text{lower groups}\end{array}}{\text{total number of students taking the test}}\right]$$

[Eq. 13.3]

The next example shows how to apply Equation 13.3 with the data in the class summary form for Item 3:

Example

How to calculate the item difficulty index for Item 3 from the data in the class summary using Equation 13.3

$$p = \left[\frac{\begin{array}{c}\text{number of students choosing the correct answer}\\ \text{for the upper} + \text{middle} + \text{lower groups}\end{array}}{\text{total number of students taking the test}}\right]$$

$$= \left[\frac{10 + 7 + 12}{34}\right] = 0.85$$

As we will discuss later, this fraction can range from 0.00 to 1.00.

Compute the Item Discrimination Index (D) (Step 6)

The **item discrimination index (*D*)** is the difference between the fraction of the upper group answering the item correctly and the fraction of the lower group answering it correctly. The discrimination index describes the extent to which a particular test item is able to differentiate the higher-scoring students from the lower-scoring students. The following equation is used to compute this index:

$$D = \left[\begin{array}{c}\text{fraction of the}\\ \text{upper group answering}\\ \text{the item correctly}\end{array}\right] - \left[\begin{array}{c}\text{fraction of the}\\ \text{lower group answering}\\ \text{the item correctly}\end{array}\right]$$

[Eq. 13.4]

Another way you will see this expressed is:

$$D = p_U - p_L$$

This index is sometimes referred to as the *net* D *index of discrimination*. Commercial test developers seldom use net *D* today; they now use a correlation coefficient as a discrimination index or other indices based on mathematical modeling of item responses. Net *D* is probably the most useful discrimination index available for use with teacher-made assessments, however.

Here is an example of how to calculate the discrimination index for one item:

Example

How to calculate the item discrimination index for Item 3 from the data in the class summary using Equation 13.4

$$D = \frac{10}{10} - \frac{7}{10} = 1.0 - 0.7 = 0.3$$

As we will discuss later, this index can range from −1.00 to +1.00.

Item Analysis of Constructed-Response and Performance Assessments

The *concepts* of item difficulty and item discrimination extend to tasks with multipoint scoring such as that obtained when you use rubrics and

FIGURE 13.3 **Item responses to the first 10 items of a 59-item test taken by a group of 34 college students.**

			Item number:									
			1	2	3	4	5	6	7	8	9	10
Upper group	Doris		A	C	B	B	C	C	B	A	B	D
	Jerry		A	C	B	B	C	C	B	A	B	D
	Robert		A	C	B	B	C	C	E	A	B	D
	Elazar		A	B	B	B	C	C	B	A	B	D
	Marya		A	C	B	B	C	C	B	A	B	D
	Anna		A	C	B	B	C	C	B	A	B	D
	Diana		A	C	B	B	C	C	B	A	B	D
	Harry		A	C	B	B	C	C	B	A	B	D
	Anthony		A	C	B	B	C	C	B	A	B	D
	Carolyn		A	C	B	B	B	C	B	A	B	D
	Key		A	C	B	B	C	C	B	A	B	D
		A	10	0	0	10	0	0	0	10	0	0
	Number	B	0	1	10	0	1	0	9	0	10	0
	choosing	C	0	9	0	—	9	10	0	0	0	0
	each	D	0	—	0	—	0	0	0	0	0	10
	option	E	—	—	—	—	—	—	1	—	—	—
	Omits		0	0	0	0	0	0	0	0	0	0
Middle group	No. right		14	12	12	13	12	13	11	11	12	12
	No. omits		0	0	0	0	0	0	0	0	0	0
Lower group	Anita		A	C	B	B	D	C	E	A	A	D
	Larry		A	C	B	B	D	C	D	A	B	D
	Charles		C	B	B	B	C	C	B	A	B	C
	Joel		A	C	B	B	C	C	E	A	B	D
	Leslie		A	C	B	B	C	C	E	A	B	B
	Alida		A	C	B	B	C	C	A	B	B	B
	Marilyn		A	C	D	B	C	C	D	C	D	A
	Wayne		A	B	A	A	C	C	B	B	C	A
	Ina		D	C	C	A	B	B	C	B	A	D
	Donald		C	B	B	B	D	C	E	C	D	D
	Key		A	C	B	B	C	C	B	A	B	D
		A	7	0	1	2	0	0	1	5	2	2
	Number	B	0	3	7	8	1	1	2	3	5	2
	choosing	C	2	7	1	—	6	9	1	2	1	1
	each	D	1	—	1	—	3	0	2	0	2	5
	option	E	—	—	—	—	—	—	4	—	—	—
	Omits		0	0	0	0	0	0	0	0	0	0

Note: This is an example of the basic data needed to do an item analysis. For the middle group you record only the number choosing the right answer and the number of omits.

rating scales. Constructed-response items and performance assessment tasks often are scored on a scale from 0 to 3, 1 to 4, or some other range of scores, instead of scoring 0 or 1. The *difficulty* of an essay question or performance task scored by a rubric or rating scale is defined as the average score. The *discrimination* of an essay question or performance task scored by a rubric or rating scale is the difference between the upper and lower group averages. There are many other ways to compute difficulty and discrimination, but we shall limit our discussion to simple ways appropriate for classroom assessment. You should monitor these values, making sure that your tasks are at an appropriate level of difficulty and that they discriminate—that is, that students who are more accomplished do indeed score better on the tasks intended to demonstrate that accomplishment.

Compute the Item Difficulty Index (p*)

The item difficulty for a constructed-response or performance item is simply the average score for that item. For example, if Item 1 was an essay item, scored on a scale of 1 to 6, and if the average score on this item was 4.2, then the difficulty of the item is 4.2.

To keep this item difficulty on the same scale as the *p*-value of Equation 13.3, we should adjust this average. This will give us a value that is between 0 and 1.00, the same as in Equation 13.3. The lowest possible score is subtracted from the average score, then divided by the possible range of scores to make the minimum value of *p** be 0.00. This difficulty index is illustrated here:

$$p^* = \frac{\begin{matrix}\text{average score}\\ \text{for}\\ \text{the item}\end{matrix} - \begin{matrix}\text{minimum}\\ \text{possible}\\ \text{item score}\end{matrix}}{\begin{matrix}\text{maximum}\\ \text{possible}\\ \text{item score}\end{matrix} - \begin{matrix}\text{minimum}\\ \text{possible}\\ \text{item score}\end{matrix}} \qquad \text{[Eq. 13.5]}$$

Here are examples of how to use this equation:

Example

Examples of applying Equation 13.5

1. Suppose the class average score on Item 1 (an essay item) was 4.2. Suppose further that the essay was scored on a scale from 1 to 6. Thus, lowest *possible* score was 1 and the highest *possible* score was 6. What is the *p** difficulty index?

$$p^* = \frac{\begin{matrix}\text{average score}\\ \text{for}\\ \text{the item}\end{matrix} - \begin{matrix}\text{minimum}\\ \text{possible}\\ \text{item score}\end{matrix}}{\begin{matrix}\text{maximum}\\ \text{possible}\\ \text{item score}\end{matrix} - \begin{matrix}\text{minimum}\\ \text{possible}\\ \text{item score}\end{matrix}}$$

$$= \frac{4.2-1}{6-1} = \frac{3.2}{5} = 0.64$$

2. Suppose on Item 2 (also an essay item) the class average was 4.2. Suppose further that it was scored on a scale from 0 to 10. Thus, the lowest *possible* score was 0 and the highest *possible* score was 10. What is the difficulty index?

$$p^* = \frac{\begin{matrix}\text{average score}\\ \text{for}\\ \text{the item}\end{matrix} - \begin{matrix}\text{minimum}\\ \text{possible}\\ \text{item score}\end{matrix}}{\begin{matrix}\text{maximum}\\ \text{possible}\\ \text{item score}\end{matrix} - \begin{matrix}\text{minimum}\\ \text{possible}\\ \text{item score}\end{matrix}}$$

$$= \frac{4.2-0}{10-0} = \frac{4.2}{10} = 0.42$$

You can see from these two examples that by taking into account the minimum possible score and the possible score range the interpretation of the average score becomes clear. In Item 1 the average score is 4.2 but the minimum possible score is 1 and possible range of marks is 1 through 6. Thus, *p** = 0.64 means that on average, students received 64% of the maximum possible score range for this item. Item 2, however, has a different interpretation. Item 2 has the same average score, 4.2. However, the minimum possible score is 0 and the possible score range for this item is from 0 to 10. The difficulty index then is 0.42, meaning that on average, students received only 42% of the maximum possible score range for this item. Thus, Item 2 is much harder for the students than is Item 1. So as you see, *you cannot fully interpret the average score of a performance item unless you know the range of the possible marks*.

Incidentally, to distinguish Equation 13.5 from the earlier Equation 13.3, we used an asterisk (*) along with *p*. This is not standard.

Compute the Item Discrimination Index (*D**)

The discrimination index for items such as constructed-response and performance items that have multipoint scoring is simply the difference between the average score on the item for the upper group and the corresponding average for the lower group. The upper and lower groups are defined in the same way as described previously; that is, based on their ranking in the total assessment. Here is an example:

Example

Suppose the upper group's average for an item is 5.3, and the lower group's corresponding average is 2.8, and the item is scored from 1 to 6. The discrimination for the item is 5.3 − 2.8 = 2.5.

To keep the item discrimination index on the same scale as the *D*-value of Equation 13.4, we adjust this difference by dividing it by the *possible* score range. This gives us a possible range for the discrimination index of between ˜1.00 and +1.00, just as with Equation 13.4. This is summarized in Equation 13.6.

$$D^* = \frac{\begin{bmatrix} \text{average score of} & & \text{average score of} \\ \text{the upper group} & - & \text{the lower group} \\ \text{on the item} & & \text{on the item} \end{bmatrix}}{\begin{bmatrix} \text{maximum} & & \text{minimum} \\ \text{possible} & - & \text{possible} \\ \text{item score} & & \text{item score} \end{bmatrix}}$$

$$= \frac{\begin{bmatrix} \text{different betweem the upper} \\ \text{and lower groups' average score} \end{bmatrix}}{[\text{range of possible scores}]}$$

[Eq. 13.6]

The next example shows how to use this equation:

Example

Example of applying Equation 13.6

Suppose the upper group's average for an item is 5.3, and the lower group's corresponding average is 2.8, and the item is scored from 1 to 6. What is the discrimination index, *D**?

$$D^* = \frac{5.3 - 2.8}{6 - 1} = \frac{2.5}{5} = 0.5$$

Because we divide by the possible item score range, we can interpret this value to mean that the difference between the average scores of the upper and lower groups for this item is 50% of the possible item score range. (This item discriminates fairly well.) As with the previously discussed discrimination index for dichotomous items (Equation 13.4), the index here can show negative values, zero, or positive values. If the value is negative, this means that the lower group scored higher on the average than the upper group. We would generally consider such a result to mean that the item is not good.

ITEM DIFFICULTY INDEX

Effect on Test Score Distribution

Shape of the Distribution The difficulty of test items affects the shape of the distribution of total test scores. Very difficult tests, containing items with *p*-values < 0.25, will tend to be positively skewed, whereas easy tests, containing items with *p*-values > 0.80, will tend to be negatively skewed. (See Figure I.7 in Appendix I for an explanation of distribution shapes.) The shapes of total score distributions for other kinds of assessments are not so easily deduced.

Average or Mean Test Score The difficulty of items affects the average or mean test score: The average test score (*M*) is equal to the sum of the difficulties of the items. The relationship is given here:

$$M = \sum p \qquad \text{[Eq. 13.7]}$$

The mean (*M*) test score is equal to the sum of the difficulty values (that is, the *p*-values) of the items comprising the test. When the assessment contains only performance or constructed-response items, the mean is simply the sum of the item means, *not* the sum of the *p** values from Equation 13.5.

Spread of Scores The spread of item difficulties and the spread of test scores are related. A test with all *p*-values clustered around 0.50 has the largest spread of test scores, whereas tests with difficulties distributed between 0.10 and 0.90 have smaller score spreads.

Item difficulties (*p*-values) are not the sole factor contributing to the spread of test scores. Another factor is the correlation (Appendix I) among the items: The higher these item intercorrelations, the larger the test's standard deviation. However, the correlations among items may be affected by the *p*-values: Items for which $p = 0.00$ or 1.00 have correlations of 0.00.

Uses of Item Difficulty Information

Figure 13.4 summarizes some of the ways in which teachers and school officials can use *p*-values and *p**-values in assessment and instruction. For the teacher, perhaps identifying concepts to be retaught and giving students feedback about their learning are the more important uses of item difficulty data. Using item information to determine curriculum strengths or to identify suspected item bias requires districtwide cooperation. Such analyses tend to be employed only with state-mandated and standardized tests because test publishers

FIGURE 13.4 Examples of ways in which item difficulty indexes can be used in testing and instruction.

Purpose	Procedure	Comments
Identifying concepts that need to be retaught	Find items with small *p*-values. These items may point to objectives needing to be retaught.	a. Poor test performance may not reflect poor teaching: Poor performance may reflect poorly written items, incorrect prior learning,or poor motivation on tests. b. A score based on several similar items is more reliable than performance on a single item.
Providing clues to possible strengths and weaknesses in school curricula	Calculate *p*-values for clusters of similar items for a school building or district. Compare these to *p*-values of the same items from the publisher's national norm group. Note areas of strength and weakness.	a. See a and b above. b. This procedure applies to standardized tests only. c. Items must correspond to local curriculum objectives and instruction. d. No published test will cover all the objectives of a school district.
Giving feedback to students	Report *p*-value of each item to student along with ID number of the items missed.	a. Such reporting is more useful for high school and college students.

make this information readily available to district offices. You may find yourself involved in interpreting state-mandated or standardized test data if you serve on school committees or if you serve in administrative positions.

ITEM DISCRIMINATION INDEX

The way you use item discrimination values should depend on the purpose of the assessment: Are you interested in absolute or relative achievement? The main purpose of **absolute achievement** assessment is to determine accurately the content or behavior each student has learned. The main purpose of **relative achievement** assessment, on the other hand, is to accurately determine the rank ordering among students with respect to the content or learning targets learned. When you are gathering information mainly about the rank order of students, you should revise or remove from the test items that do not contribute information about ordering students or that provide inconsistent, confusing information about this ordering.

Suppose you wanted to order a class of students from high to low using a 30-item unit test. Suppose, further, that when doing an item analysis, you divide the class in half based on the total test score (as usual, higher scorers in the upper group, lower scorers in the lower group). Finally, suppose that for one of the items, you discover that all of the lower-group students answered the item correctly, and the entire upper group answered it incorrectly. In this case, the item difficulty index is $p = 0.50$, but the item discrimination index is $D = 0 - 1.00 = -1.00$. This negatively discriminating item is poor because it works in the opposite way from most of the other items. That is, high-scoring students answer it incorrectly and low-scoring students answer it correctly. If you were to put such negatively discriminating items on a test, they would work to arrange students in an order inconsistent with the arrangement resulting from the positively discriminating items on the assessment.

Only the discrimination index is able to detect the type of malfunctioning item just described. The difficulty index gives the proportion of the class that answers an item correctly, but it does not indicate whether more higher- or lower-scoring students answered correctly. *For this reason, you should give more weight to an item's discrimination index than its difficulty index when deciding whether the item should appear on a test*.

Numerical Limits of D

For each item, the possible net D range is from −1 to +1. If all the discriminations made by an item were correct discriminations (everyone in the upper group answers the item right, whereas everyone in the lower group answers it wrong), net D would equal +1. Such an item is said to be a perfect **positively discriminating item**. If the number of correct discriminations equals the number of incorrect discriminations (an equal number of upper- and lower-group students answer the item right), then $D = 0$. Such an item is said to be a

nondiscriminating item. Finally, if all discriminations were incorrect (everyone in the upper group answers the item wrong while everyone in the lower group answers it right), the D would equal −1. Such an item is said to be a perfect **negatively discriminating item**.

The values +1 and −1 are seldom obtained in practice. $D = 0$ is obtained most often for very easy or very hard items. The values 1, 0, and +1 serve as benchmarks when interpreting D.

Score Reliability and Item Discrimination Power

If none of the items discriminated ($D = 0$ for all items), everyone would be bunched together. If individual items can't distinguish students, then the collection of items comprising the test won't be able to do so, either. The larger the test's average level of item discrimination, the more diverse the scores will be. A more reliable assessment will be made up of tasks with high, positive discrimination indices. Thus, if the primary purpose of using an assessment is to interpret differences in achievement among students, the assessment procedure must include tasks with high discriminating power.

Interpret a negative value of D as a warning that you should carefully study the item and either revise or eliminate it. If you cannot find a technical flaw in the item, it might be that students in the upper-scoring group learned the material either incompletely or entirely incorrectly. Barring any rational explanation to the contrary, all of your assessment's items should be positively discriminating; otherwise, the total score on the assessment won't provide usable information.

IMPROVING MULTIPLE-CHOICE ITEM QUALITY

Poorly Functioning Distractors

Response Patterns for Distractors The main purpose of the distractors or foils in a multiple-choice item is to appear plausible to those students lacking sufficient knowledge to choose the correct answer. Item analysis data of the type summarized by the class record shown in Figure 13.3 can be used to find out which item distractors are not meeting this purpose and are therefore **poorly functioning distractors**. The general rule is this: *Every distractor should have at least one lower-group student choosing it, and more lower-group students than upper-group students should choose it*.

Because of fluctuations in responses from one small group of students to another, use the rules of thumb carefully. The following data, from the item presented previously on the item file card example, illustrates these points. Each distractor (B, C, and D) was chosen by at least one lower-group person; no student in the upper group chose a distractor.

Example

Example of item analysis data showing an appropriate pattern of responses to distractors

Alternative	Upper group	Lower group
*A	10	7
B	0	1
C	0	1
D	0	1

The rationale for the general rule is as follows: Students scoring lowest on the test are, on the whole, least able (in a relative, not absolute, sense) regarding the performance being assessed. If they are not, then the test on which they scored lowest must lack validity. For every item, it is among these lower-scoring students that you should expect to find incorrect alternative (distractors) chosen. Thus, if an item is working properly, one or more lower-scoring students should choose each distractor, and more lower-scoring than upper-scoring students should choose distractors.

Notice that not every lower-scoring person lacks knowledge about every item: In the preceding example, 7 out of 10 lower-scoring persons knew the answer. Neither is it the case that every higher-scoring person always chooses the correct answer (see, for example, Items 2, 5, and 7 in Figure 13.3).

If no student in the lower group chooses a particular distractor, the distractor may be functioning poorly. Here is an example of a response pattern that shows that Distractor B may not be functioning properly. This example is for Item 1 of Figure 13.3.

Example

Example of item analysis response pattern showing that Alternative B should be checked to see if it is functioning poorly

Alternative	*Upper group*	*Lower group*
*A	10	7
B	0	0
C	0	2
D	0	1

You should review the item and speculate why this occurred. Perhaps the particular alternative contains one of the technical flaws described in Chapter 9. If all students recognize a particular option as obviously incorrect, then you will want either to eliminate the alternative entirely (thus reducing the number of options in the item), substitute an entirely new alternative, or revise the existing alternative.

Subject Matter Has Precedence It isn't always true that an alternative is flawed if no one in the lower group chooses it. Here's where your knowledge of the subject matter, of the students, and of the instruction students received prior to taking the assessment come into play: Perhaps in this year's group, even the lowest-scoring students have enough knowledge to eliminate a particular distractor, yet they do not have enough knowledge to select the correct answer. Perhaps in other groups a concept will not be learned as well, and this particular distractor will be plausible. Eliminating the alternative would prevent you from identifying those few individuals who lack this learning. In other words, use your own expertise along with the data to decide whether to eliminate a distractor that isn't working.

Finally, note that even though it seems reasonable to *expect* a larger number of lower-scoring students than higher-scoring students to choose a particular distractor, this may not always happen. Technical flaws may cause higher-scoring students to be deceived, such as when they know a great deal about the subject and thus are able to give a plausible reason why an unkeyed alternative is at least as correct as the keyed one. In such cases, the alternative definitely should be revised.

But sometimes there is neither a technical flaw nor a subject-matter deficiency in an incorrect alternative, yet higher-scoring students choose it in greater numbers than lower-scoring ones. In these cases, students may have incomplete or wrong learning. Examples of such two items were shown in Chapter 9.

Ambiguous Alternatives

Student responses can provide leads to **ambiguous alternatives**. In this context, alternatives are ambiguous if *upper-group students* are unable to distinguish between the keyed answer and one or more of the distractors. When this happens, the upper group tends to choose a distractor with about the same frequency as the keyed response, as illustrated in the following example:

Example

Example of an upper-group distractor response pattern showing ambiguous alternatives

On which river is the city of Pittsburgh, Pennsylvania, located?	*Upper group*	*Lower group*
A Delaware River	0	3
B Ohio River	5	3
C Monongahela River	4	1
D Susquehanna River	1	3

The confluence of the Allegheny and Monongahela Rivers forms the Ohio River at Pittsburgh. The upper group in the example chose B and C with approximately equal frequency, thus reflecting the students' ambiguity in selecting only one of these two alternatives as a correct answer. This item should be rewritten so that only one answer is clearly correct or best.

You might notice that very often the lower group is equally divided among two or more alternatives. This is usually *not* an indication that you must revise the item. Rather, it means that students with less knowledge will find many alternatives equally plausible, and so the task becomes an ambiguous one for them. The cause of these students' ambiguity is likely to be insufficient knowledge.

Before concluding that you need to revise an item, however, study the item in relation to the students taking the test and judge whether the ambiguity stems from the students' lack of knowledge rather than from a poorly written item. Consider the next example, which shows how incomplete learning may produce a response pattern that gives the appearance of ambiguous alternatives.

Example

Example of how incomplete learning may result in a response pattern that gives the appearance of ambiguous alternatives

$3 + 5 \times 2 = ?$

	Upper group	Lower group
A 10	0	2
*B 13	5	3
C 16	5	3
D 30	0	2

This item requires applying arithmetic operations in a certain order: multiplication first, then addition. Option B reflects this order, whereas Option C is the answer obtained by adding first and then multiplying. Apparently half the upper group followed this erroneous procedure and chose C. The item is not technically flawed, but the responses indicate to the teacher that a number of students need to learn this principle. The entire group's responses to this item should be checked, of course.

Miskeyed Items

You may have **miskeyed** an item if a larger number of upper-group students select a particular wrong response. When this happens, check to be sure that the answer key is correct. Look at this example:

Example

Example of an upper-group distractor response pattern showing a possible miskeyed item

Who was the fourth president of the United States?	*Upper group*	*Lower group*
A John Quincy Adams	0	3
B Thomas Jefferson	1	2
C James Madison	9	3
*D James Monroe	0	2

In the example, C is the correct answer, but the teacher inadvertently used Alternative D as the answer key. The response pattern in the figure is typical of such an item.

Again, be sure to check the item content. The numbers from the item analysis only warn of possible miskeying—perhaps there is no miskeying and the upper group simply lacks the required knowledge.

Random Guessing

Students may be guessing randomly if many of the alternatives are equally plausible to the upper-scoring group. If the upper-group students guess randomly, each option tends to be chosen an approximately equal number of times, as illustrated in this example:

Example

Example of an upper-group distractor response pattern showing possible random guess or confusion

In what year did the United States enter World War I?	*Upper group*	*Lower group*
A 1913	2	3
B 1915	2	2
C 1916	3	3
*D 1917	3	2

Remember to look at the pattern of responses of the upper group, *not* the lower group, to find items on which many students may be guessing. Guessing among the most knowledgeable students may signal widespread confusion in the class. Lower-scoring students may in fact be guessing on the more difficult items, too, but this indicates you need to reteach them rather than simply revise the test item. Random guessing adds errors of measurement to the scores, thereby reducing reliability and validity.

SELECTING TEST ITEMS

Another Purpose for Item Analysis

Most teachers who use item analysis procedures do so for one or more of the following reasons: (a) to check whether the items are functioning as intended, (b) to give students feedback on their assessment performance, (c) to acquire feedback for themselves about students' difficulties, (d) to identify areas of the curriculum that may need improvement, and (e) to obtain objective data that signal the need for revising their items. You can also use item analysis for selecting some items and culling others from a pool of items.

Purpose of Assessment Helps Select Items

No statistical item selection rule is helpful if it is inconsistent with your purpose for conducting assessments. Further, any procedure you use for selecting some items over others changes the definition of the domain of performance. Those performances represented by items you eliminate are never assessed.

Relative Versus Absolute Student Attainment Careful selection of items results in shorter, more efficient, and more reliable assessments. In the classroom, *statistically based item selection* seems to apply most when you are concerned primarily with students' relative achievement rather than their absolute achievement. You focus on assessing relative achievement when your priority is to rank students with respect to what they have learned. You focus on assessing absolute achievement when your priority is to determine the precise content (or performance) each student has learned.

As an example, suppose you wanted to assess students' learning of the 100 simple addition facts typically taught in first and second grades. If you want to know only the relative achievement of the students (which student knows the most, next most, and so on), you could use a relatively short test, made up of only addition facts that best discriminate among the students. This test would probably contain mostly the middle and upper parts of the addition table. Addition facts that almost everyone knows (such as $1 + 1 = 2$) would not be included on such a test because these items would not discriminate ($D = 0$) and thus would not provide information to rank the students. However, excluding certain addition facts from the assessment because they do not discriminate well means that you will be unable to observe a student's performance on all 100 addition facts.

On the other hand, suppose your purpose for assessment is to identify the particular addition combinations with which a student has difficulty. In this case, finding out the absolute level of achievement would be your main assessment focus. You may find it necessary to use a longer, less efficient test (or several shorter ones), perhaps assessing all 100 facts.

Absolute, rather than relative, achievement is more important for diagnostic assessments intended to identify such things as whether a student has acquired particular reading skills, learned a certain percentage of facts in some specified domain, or has the ability to solve certain types of problems. Relative achievement is more important when you are assessing a student's general educational development in a subject area.

Complete Versus Partial Ordering For some educational decisions, you may need to accurately rank all students using their test performance, called a **complete ordering of students**. On the other hand, you may only want to separate students into five ordered categories so you can assign grades (A, B, C, D, and F). In so doing, you may not wish to make precise distinctions among the students within each category. Similarly, you may wish to divide the class into two groups, such as mastery/nonmastery or faster/slower readers. We say there is **partial ordering** when the categories themselves are ordered, but there is *no ordering of individuals within a category*. Categorizing students by their grades, or into fail-pass groups, are examples of partial ordering.

When you focus your assessment on either partial or complete ordering, it is inefficient to include items that do not contribute to ordering and distinguishing students. Such items therefore are culled from the pool. To cull, you try out items with students before creating the final version of the test (or you use items from past administrations of the test for which you have data). Calculate item statistics (p and D). Select and assemble into the final test those items with high, positive discrimination indices. Select items with p-values (difficulty) at each level of performance where you wish to have information (e.g., A through F). C students, for example, should not simply be C students because they got right, partly by chance, a portion of the items that A students are expected to get. C students should be in that category because they scored correctly on items at that level of difficulty.

Realities, Content Coverage, and Compromise

In practice, you must include test items with less than ideal statistical properties so a test can match its blueprint. Actual assessment construction tends to be a compromise between considerations of subject-matter coverage and psychometric properties. The general principle is: *Select the best available items that cover the important areas of content as defined by the blueprint, even though the discrimination and difficulty indices of these items have values that are less than ideal.*

Rules of Thumb for Selecting Test Items

Figure 13.5 summarizes guidelines for selecting items for classroom tests, keeping in mind our discussion of the differences between building a test to measure relative achievement and building one to measure absolute achievement. Note that coverage of content and learning targets has primacy over statistical indexes when selecting test items by the procedures recommended here. The guidelines shown in Figure 13.5 require you to understand whether the prospective test should assess only one ability or a combination of several abilities. A **homogeneous test** will measure one ability, whereas a **heterogeneous test** will assess a combination of abilities. If your test contains some items for which students can get the right answer by random guessing (such as with multiple-choice items), then the items you select should be approximately 5% easier than shown in the figure.

In choosing between two items assessing the same learning target for a test of relative achievement, good item discrimination takes precedence over obtaining the ideal item difficulty level. That is, if two items assess the same learning target and are of approximately the same difficulty level, use the one that discriminates better.

When you design a criterion-referenced classroom test, item statistics play a lesser role for selecting and culling items. You should still calculate item statistics to obtain data on how the items might be improved, however. Items exhibiting zero or negative discrimination frequently contain technical flaws that you may not notice unless you do an item analysis. You should also make sure that the item difficulties cover the range of expected performance.

FIGURE 13.5 **Guidelines for selecting items.**

	Relative achievement is the focus		
	Complete ordering	**Partial ordering (two groups)**	**Absolute achievement is the focus**
General concerns	Ranking all the pupils in terms of their relative attainment in a subject area.	Dividing pupils into two groups on the basis of their relative attainment. Pupils within each group will be treated alike.	Assess the absolute status (achievement) of the pupil with respect to a well-defined domain of instructionally relevant tasks.
Specific focus of test	Seek to accurately describe differences in relative achievement between individual pupils.	Seek to accurately classify persons into two categories.	Seek to accurately estimate the percentage of the domain each pupil can perform successfully.
Attention to the test's blueprint	Be sure that items cover all important topics and objectives within the blueprint.	Be sure that items cover all important topics and objectives within the blueprint.	Be sure items are a representative, random sample from the defined domain that the blueprint operationalizes.
How the difficulty index (*p*) is used	Within each topical area of the blueprint, select those items with: (1) *p* between 0.16 and 0.84 if performance on the test represents a single ability. (2) *p* between 0.40 and 0.60 if performance on the test represents several different abilities. *Note:* Items should be easier than described above if guessing is a factor.	Within each topical area of the blueprint, select those items with *p*-values slightly larger than the percentage of persons to be classified in the upper group (e.g., if the class is to be divided in half [0.50] then items with *p*-value of about 0.60 should be selected; if the division is lower 75% vs. upper 25%, items should have $p = 0.35$ [approximately]). *Note:* The above suggestion assumes the test measures a single ability.	Don't select items on the basis of their *p*-values, but study each *p* to see if it is signaling a poorly written item. Make sure there is a sufficient number of items with *p* values at each level of performance.
How the discrimination index (*D*) is used	Within each topical area of the blueprint, select items with *D* greater than or equal to +0.30.	Within each topical area of the blueprint, select items with *D* greater than or equal to +0.30.	All items should have *D* greater than or equal to 0.00. Unless there is a rational explanation to the contrary, revise those items not possessing this property.

USING COMPUTERS AS AN AID IN TESTING

Using Computers for Test Assembly and Item Analyses

In some schools, students' responses on special answer sheets can be scanned directly into a program that does all the item analyses illustrated in this chapter. In other schools, the scanner creates a computer file, but you must use your own program to analyze the data. You can duplicate much of the analysis done by specialized programs using a standard spreadsheet program that comes with office suite programs. Assessment Systems Corporation has free software for analyzing classroom tests (limit 50 items and 50 examinees). It is called CITAS (Classic Item and Test Analysis Spreadsheet) and is available at www.assess.com.

Vendors have also created software that allows banking or storing test items in a computer file (i.e., both the item's text and graphics). The software then allows you to select items from the bank, assemble tests, and print them for duplication. Some software permits tests to be administered via intranet or Internet. Other software products offer even more organization: You can align your assignments with your state's standards and school's curriculum objectives, then compare each student's progress against these standards and objectives.

Software, hardware, and related products vary greatly, not only in their quality, cost, and user friendliness but also in how well they match your teaching and school's instructional goals. Some programs can be run right out of the box, whereas others require considerable training. We are not able to review the products here. You can visit the *T.H.E. Journal* Website (http://www.thejournal.com) or Websites of firms that produce assessment and item-analysis software (e.g., Assessment Systems Corporation at http://www.assess.com).

Using Technology to Make Tests More Accessible to Students With Disabilities

Many nonessential elements of testing can be altered or enhanced with computer applications that make tests more accessible to students with disabilities. These changes can be as simple as enlarging the font on a test, using a word processing program. They can be as complex as using augmentative communications systems for students who cannot speak. Dictionaries, thesauruses, and grammar checkers can help students prepare written test answers. Technology can assist a teacher in calculating readability of the text in assessments. The Internet can be a source of images to help make test items readable.

Teacher judgment is required to decide what accommodations are appropriate for particular students and particular assessments. For example, enlarging the font on a reading test would probably not change the construct being measured, merely make it easier for the student to read the passages. Changing the readability level of the passages, however, would change the construct being measured.

Technological solutions to problems of accessibility are changing at a fast pace. Salend (2009) organized various currently available technology solutions according to the principles of universal design. See Appendix E for this helpful list. The term *universal design* refers to the concept of preparing assessments to maximize accessibility for all students. It is a term that began in the field of architecture, and is more often discussed in terms of large-scale assessments. For this reason, we discuss universal design in more detail in Chapter 17. For present purposes, however, our point is that there are many ways to make classroom tests more accessible, and these have greatly expanded with the availability of computer applications.

CONCLUSION

In this chapter, we discussed preparing students for assessment. We also discussed how to use item analysis to assist you in maximizing the quality of your tests. Both of these are important aspects of handling assessments that sometimes do not get the thoughtful planning they deserve. We turn next to grading, which may have the opposite problem: Teachers and students sometimes devote more time and energy to grades than is probably good for them. We hope the next chapter will help you approach grading thoughtfully, as the classroom-summative part of a balanced assessment system that also includes your classroom formative assessments and external summative assessments like state tests.

EXERCISES

1. The following statements are thoughts that students might have during an assessment situation. Read each statement and decide whether it is a task-relevant (TR) thought or a task-irrelevant (TI) thought.
   a. "I have to be very careful in answering this problem. My teacher takes points off for computational errors."
   b. "I am really dumb. I just can't do it!"
   c. "If I don't pass this test, Dad will kill me!"
   d. "I know I don't know the answer to this question. It's no use trying to fool Mr. Jones. He'll just think I'm dumber than I am."
   e. "Oops! I forgot to study the material this question is asking. Oh well, I'd better write something down. I usually am able to get a few points from Mr. Jones!"
2. Explain the meaning of each of the following values of *D*.
   a. +1.00
   b. +0.50
   c. 0.00
   d. –0.50
   e. −1.00
3. Figure 13.6 shows a summary of item analysis data for five multiple-choice items for a class of 30 students. There are 11 students in the upper group and 11 students in the lower group. The keyed answer to each item is marked with an asterisk. For each item, calculate the difficulty index (*p*) and the discrimination index (*D*), then decide whether the item has poor distractors, is possibly miskeyed, the upper group is possibly guessing, or two options seem to be ambiguous.
4. The following questions refer to your analysis of the item data in Exercise 3.
   a. Which item is a negative discriminator?
   b. Which item is the easiest?
   c. Which item is the most difficult?
   d. For which items do more upper-group students than lower-group students choose a distractor?
   e. Which item has the highest discrimination index?
   f. Which item has the lowest discrimination index?
   g. What is the average (mean) score on this five-item test for the 30 students who took it?

FIGURE 13.6 **Item analysis summary for use with Exercise 4.**

		Options							
Item number	**Groups**	**A**	**B**	**C**	**D**	**Faulty distractors**	**Miskeying**	**Ambiguous**	**Guessing**
1.	Upper	0	2	*9	0	——	——	——	——
	Middle			*5					
	Lower	1	2	*4	4		*p* = ——	*D* = ——	
2.	Upper	2	*7	0	2	——	——	——	——
	Middle		*4						
	Lower	0	*9	1	1		*p* = ——	*D* = ——	
3.	Upper	9	*1	1	0	——	——	——	——
	Middle		*1						
	Lower	6	*2	2	1		*p* = ——	*D* = ——	
4.	Upper	*5	5	0	1	——	——	——	——
	Middle	*8							
	Lower	*3	3	3	2		*p* = ——	*D* = ——	
5.	Upper	3	2	3	*3	——	——	——	——
	Middle				*4				
	Lower	3	2	3	*3		*p* = ——	*D* = ——	

Evaluating and Grading Student Progress

KEY CONCEPTS

1. The main purpose of grading is to communicate information about student achievement.
2. Report cards are one of several means of reporting student progress.
3. A criterion-referenced grading model matches the typical standards-based or objectives-based approach to teaching.
4. Choose and weight components for grading according to your assessment plan. Grading creates a measurement scale that—like any scale—should be valid and reliable.
5. There are norm- and criterion-referenced methods for combining scores into one summary achievement grade. You should choose the one appropriate to your situation. Criterion-referenced grading is recommended.

IMPORTANT TERMS

assessment variables (evaluation variables)
borderline cases
continuous assessment
criterion-referenced grading framework (absolute standards)
fixed-percentage method for grading
gradebook program
grading
grading for summative purposes
grading on a curve
grading variables
logic rule method for grading
median score method
minimum attainment method
multiple marking system
narrative report
norm-referenced grading framework (relative standards)
permanent record
quality-level method for grading (content-based method, rubric method)
report card
reporting variables
self-referenced grading framework (growth standards)
SS-score method for making composites
stakeholders

From Chapter 12 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

standard deviation method of grading

student progress reporting method (checklist, letter grades, letter to parents, narrative reports, numbers, parent-teacher conferences, percentages, pupil-teacher conferences, rating scale, two-category)

total points method for grading

THE MEANINGS AND PURPOSES OF GRADES

What are Your Attitudes toward Marks and Grades?

Before starting this chapter, consider how you feel about assigning grades and marks. Read each of the statements in Figure 14.1. Next to each one, check A if you agree, D if you disagree, and U if you are undecided. Compare your answers with those of your classmates and your instructor. Keep these attitudes in mind as you study this chapter and think about how to apply the concepts to your own teaching. Revisit your answers after you study this chapter. How many answers did you change?

Continuous Assessment and Grading

Formative Assessment **Continuous assessment** is the daily process by which you gather information about students' progress in achieving the curriculum's learning targets (Nitko, 1995). Continuous assessment has both formative and summative aspects. You use formative continuous assessment to make decisions about daily lesson planning and how well your day's lesson is going. You do not formally record formative evaluations on a report card or a permanent record card. Many formative evaluations are reported directly to the student.

Summative Assessment This chapter emphasizes how to use grades to report your summative continuous assessments of students' achievement of the curriculum's major learning targets. **Grading** (or marking) refers primarily to the process of using a system of symbols (usually letters) for reporting various types of student progress. **Grading for summative purposes** lets you provide yourself, other teachers, school officials, students, parents, postsecondary educational institutions, and potential employers with a report about how well students have achieved the curriculum learning targets. You usually are required to report students' grades several times a year to parents or guardians. The report covers several weeks of school, called a marking period; this is often each

FIGURE 14.1 **What are your feelings about marks and grades (A = agree, U = undecided, D = disagree)?**

A U D

1. There are justifiable reasons why the marks of some teachers, courses, and departments average consistently higher than others. __ __ __
2. Academic marks should be based more on achievement status than on growth or progress. __ __ __
3. Students' academic marks should be determined solely by their academic achievements and not by attendance, citizenship, effort, and attitudes. __ __ __
4. Schools that use marks should adopt and enforce a clearly defined institutional marking policy. __ __ __
5. In the absence of an institutional marking policy, marks should not be used in determining students' eligibility for academic courses and programs. __ __ __

A U D

6. Absolute standards are more desirable than relative standards in evaluating and marking students in academic areas. __ __ __
7. In the absence of an institutional marking policy, marks should not be used in determining eligibility for athletics and other extraclass activities. __ __ __
8. "Pass/fail" or "credit/no credit" are more desirable than marking systems with three or more categories for academic classes. __ __ __
9. Allowing students to contract for their own marks is preferable to marking on a relative basis. __ __ __
10. Teachers should attempt to evaluate and mark students in such areas as interests, attitudes, and motivation. __ __ __

quarter of the academic year. The grades you give students are reported to the school administration on a permanent record card or folder. In the later years of schooling, they become part of the student's transcript. The school reports grades to students and parents through various means such as report cards, conferences, or letters.

Validity Is Required Grades serving official summative evaluation purposes must be based on formal, continuous assessments that are aligned with your school's standards, official curriculum's learning targets, and educational psychology. As one fourth-grade teacher said:

> I don't know how other teachers feel, but anytime I send out an official report with my name on it, it is the equivalent of a legal document. The information in that report declares itself to be the best and latest educational information on a child. This may sound overly dramatic, but parents are expecting that report to tell them about an important chunk of their child's life. It is supposed to be true, and it is official. (Cited in Azwell & Schmar, 1995, pp. 7–9)

Because many **stakeholders** will use your summative grades for many different purposes, the grades must be validly prepared and based on high-quality assessments. Assessments contributing to grades come from several sources: curriculum materials, quizzes and tests, performance tasks you create, projects and other long-term tasks, products students produce, portfolios you and your students assemble, and assessments set by groups of teachers working together.

It seems unfair to base a student's final grade on a single examination (assessment). This "big bang" approach to evaluation ignores several important factors about assessing students: (a) Only a limited amount of time is available during one teaching period for assessing; (b) in a limited time, only a small sample of tasks can be administered to students; (c) students may know much more than what appeared on the "one shot" assessment; (d) students' illness or family problems can interfere with their ability to demonstrate the required achievement; (e) students can demonstrate their achievement in several ways other than the one way you decided to assess it; and (f) some important learning targets are best assessed through longer-term projects, papers, or out-of-school assignments.

Why Teachers Dislike Grading Grading for many teachers is one of the most difficult and troublesome aspects of teaching. Teachers are usually much more comfortable in their role as advocates for their students than as judges or evaluators. In spite of teachers' dislike of grading, it is a required part of the job. This is one reason why you need to learn how to grade students as validly as possible.

How People Perceive and Use Grades

Figure 14.2 gives examples of information frequently found on formal student progress reports

FIGURE 14.2 **Examples of the types of information found on report cards and the types of decisions made from that information.**

	Decisions that can be made			
Information in report	**Selection**	**Placement, remediation**	**Guidance, counseling**	**Course improvement**
1. Content or objectives learned	Promotion, probation, graduation, admissions	Selecting courses to take, remedial help needed	Selecting next courses to take, additional schooling needed, career-related choices	Deciding where instruction can be improved
2. Comparison of performance in different subjects	Admission	Selecting advanced and/or remedial courses	Determining pattern of a pupil's strengths and weaknesses	Identifying areas that are strong points of school
3. Performance relative to other people	Scholarships, prizes, admission	Estimating likely success, eligibility for special programs	Estimating likely success in certain areas	
4. Social behavior		Matching personal characteristics to course and teacher placement	Determining need for adjustment, likes, dislikes, ability to get along with others	Identifying problems with a course or with a teacher

and various kinds of decisions that may be based on such information. Different persons will use grades in different ways. Figure 14.3 shows several different types of stakeholders and the ways they use grades. This figure illustrates that grades have serious meaning beyond your classroom. The grades you assign must be clear to judge whether any of these uses are valid.

Although assessment specialists generally recommend that you keep the meaning of grades clear by basing them only on a student's achievement of your course's learning targets, we know that many teachers do not follow this advice (Brookhart, 1991; Stiggins, Frisbie, & Griswold, 1989; Waltman & Frisbie, 1994). Brookhart states the issue clearly:

> The adjustments teachers make to compensate for grade use and misuse, however, are not uniform and are not necessarily valid either. A hodgepodge grade of attitude, effort, and achievement, created in an attempt to provide positive feedback to the student, is not the answer. Such a hodgepodge grade also falls down under a validity check; it does not possess the characteristic of interpretability. What teachers seem to intend when they add nonachievement factors to grades is to mitigate negative social consequences, but grades are not the appropriate tool for social engineering. Teachers' intuition about social consequences, however, is useful because it points us to the other half of the validity issue: what happens when grades are used for decisions and actions. (p. 36)

FIGURE 14.3 **Various uses to which grades are put by different stakeholders.**

	Stakeholder likely to use the grades in the way indicated						
Usage for grades	Student	Parents	Teacher	Guidance counselor	School administrators	Postsecondary educational institutions	Employers
1. Reaffirm what is already known about classroom achievement	✓		✓				
2. Document educational progress and course completion	✓	✓	✓	✓	✓	✓	✓
3. Obtain extrinsic rewards/ punishments	✓	✓					
4. Obtain social attention or teacher attention	✓						
5. Request new educational placement		✓	✓	✓	✓		
6. Judge a teacher's competence or fairness		✓			✓		
7. Indicate school problems for a student		✓	✓	✓	✓		
8. Support vocational or career guidance explorations	✓	✓		✓			
9. Limit or exclude student's participation in extracurricular activities		✓	✓		✓		
10. Promote or retain			✓		✓		
11. Grant graduation/diploma					✓		
12. Determine whether student has necessary prerequisite for a higher-level course			✓	✓	✓		
13. Select for postsecondary education						✓	
14. Decide whether an individual has basic skills needed for a particular job							✓

Parents' Versus Teachers' Understanding

Communicating to parents is especially challenging. Some research shows that parents' and teachers' understanding of what report card grades mean are often far apart (Waltman & Frisbie, 1994). For example, parents may see grades as reflecting pure achievement. Or they may interpret the grading scale differently than teachers do, for example thinking of a C as "average" when most teachers' average grade is about a B.

Grades Communicate Your Values

For the teacher, grades communicate more than achievement information about a student. The grades you assign communicate your (and your school's) values. If obedience to your classroom rules is rewarded by an A or "performing satisfactorily" in *reading,* but "fooling around" during class means the *reading grade* is lowered, in spite of successful reading performance, you have communicated that obedience is valued more than reading well. The teacher who gives an unsatisfactory grade to the student whose academic performance is satisfactory and then says, "I warned you about passing notes during class!" is perhaps communicating vindictiveness. You may value both social behavior (e.g., conformity) and achievement, but if the grade you report intertwines the two, you are communicating poorly and are encouraging confusion. To clarify matters, you must separate your evaluations of achievement from your evaluations of noncognitive student characteristics.

Criticisms of Grades and Marks

Educators have voiced a number of criticisms of grades over the years. You need to be aware of these criticisms to explain the rationale for your own grading policy to parents and other educators. Many of these criticisms can be summarized under the four headings in Figure 14.4 (Ebel, 1974).

FIGURE 14.4 **Commonly expressed criticisms of grades.**

A. Grades are essentially meaningless.

1. There is great diversity among institutions and teachers in grading practices.
2. Many schools lack definite grading policies.
3. A single symbol cannot possibly report adequately the complex details of an educational achievement.
4. Teachers are often casual or even careless in grading.
5. Grades are frequently used to punish or to enforce discipline rather than to report achievement accurately.

B. Grades are educationally unimportant.

6. Grades are only symbols.
7. The most important outcomes are intangible and hence cannot be assessed or graded.
8. A teacher's grades are less important to pupils than their own self-evaluations.
9. Grades do not predict later achievement correctly.
10. What should be evaluated is the educational program, not the pupils.

C. Grades are unnecessary.

11. Grades are ineffective motivators of real achievement in education.
12. When students learn mastery, as they should, no differential levels of achievement remain to be graded.
13. Grades have persisted in schools mainly because teachers cling to traditional practices.

D. Grades are harmful.

14. Low grades may discourage the less able pupils from efforts to learn.
15. Grading makes failure inevitable for some pupils.
16. Parents sometimes punish pupils for low grades, and reward high grades inappropriately.
17. Grades set universal standards for all pupils despite their great individual differences.
18. Grading emphasizes common goals for all pupils and discourages individuality in learning.
19. Grading rewards conformity and penalizes creativity.
20. Grading fosters competition rather than cooperation.
21. Pressure to get high grades leads some pupils to cheat.
22. Grading is more compatible with subject-centered education than with humanistic, child-centered education.

STUDENT PROGRESS REPORTING METHODS

Student progress reporting methods are ways that schools communicate to students and parents, as well as ways of keeping records of students' achievement. Figure 14.5 summarizes the advantages and disadvantages of different methods. Your school district may use more than one method of reporting student progress because different methods may serve different purposes and different audiences.

Teachers use some methods of reporting student progress more frequently at certain grade levels. Letter grades are used with high frequency in the upper elementary, junior high, and senior high school levels. Parent-teacher conferences do not occur often in junior and senior high schools.

FIGURE 14.5 **Advantages and disadvantages of some commonly used methods of reporting student progress.**

Name	Type of code used	Advantages	Disadvantages
Letter grades	A, B, C, etc., also "+" and "–" may be added.	a. Administratively easy to use b. Believed to be easy to interpret c. Concisely summarize overall performance	a. Meaning of a grade varies widely with subject, teacher, school b. Do not describe strengths and weaknesses c. Kindergarten and primary school children may feel defeated by them
Number or percentage grade	Integers (5, 4, 3 . . .) or percentages (99, 98, . . .)	a. Same as points a, b, and c above b. More continuous than letter grades c. May be used along with letter grades	a. Same as points a, b, and c above b. Meaning not immediately apparent unless explanation accompanies them
Standards-based grade	Advanced, Proficient, Basic, Below Basic, or similar	a. Requires standards-referenced grading methods b. Often used with fine-grained reporting categories	a. May not match state test results b. Difficult to adapt for different levels of learners
Two-category grade	Pass-fail, satisfactory-unsatisfactory, credit-entry	a. Less devastating to younger students b. Can encourage older students to take courses normally neglected because of fear of lowered GPA	a. Less reliable than more continuous system b. Does not communicate enough information about pupil's performance for others to judge progress
Checklist and rating scales	Checks (✓) next to objectives mastered or numerical ratings of degree of mastery	a. Give the details of what the pupil achieved b. May be combined with letter grades or with group-referenced data	a. May become too detailed for parents to comprehend b. Administratively cumbersome for record keeping
Narrative report	None, but may refer to one or more of the above; however, usually does not refer to grades	a. Allows teacher the opportunity to describe a student's educational development b. Shows a student's progress in terms of standards, indicators of achievement, learning targets, or a continuum of educational growth	a. Very time-consuming b. Requires excellent writing skill and effective communication skills on the teacher's part c. May require translation into language read by parents, with possible loss of meaning in the translation

FIGURE 14.5 (*Continued*)

Name	Type of code used	Advantages	Disadvantages
		c. Provides opportunity to open dialogue and other types of communication with parents and students	d. Parents who are not skilled readers may misunderstand it or may be put off e. Parents may be overwhelmed and not respond f. Often modified to include checklist-like list of indicators with short teacher comments
Pupil-teacher conference	Usually none, but any of the above may be discussed	a. Offers opportunity to discuss progress personally b. Can be an ongoing process that is integrated into instruction	a. Teacher needs skill in offering positive as well as negative comments b. Can be time-consuming c. Can be threatening to some pupils d. Doesn't offer the institution the kind of summary record desired
Parent-teacher conference	None, but often one or more of the above may be discussed	a. Allows parents and teachers to discuss concerns and clarify misunderstandings b. Teachers can show samples of students' work and explain basis for judgments made c. May lead to improved home-school relations	a. Time-consuming b. Requires teacher to prepare ahead of time c. May provoke too much anxiety for some teachers and parents d. Inadequate means of reporting large amounts of information e. May be inconvenient for parent to attend
Letter to parents	None, but may refer to one or more of the above	a. Useful supplement to other progress-reporting methods	a. Short letters inadequately communicate pupil progress b. Requires exceptional writing skill and much teacher time

Often schools use combinations of methods on the same report card. For example, letter grades may report students' subject-matter achievement; rating scales may report the students' attitudes and deportment. A parent-teacher conference may convey information on achievement, effort, attitudes, and behavior. Schools may use a combination of nearly all methods.

Conflicts may arise between methods. School administrators need a concise summary of each student's progress for accountability and record keeping. Parents and teachers may need slightly more detailed explanations of the content taught, the standards mastered, and how a student's educational development compares with members of a peer group. The most detailed methods of reporting identified in Figure 14.5 are the checklist and the narrative.

Checklists

A checklist contains a list of many specific behaviors; a teacher checks off or rates each behavior as a student performs it during the year. We discussed how to construct checklists in Chapter 12.

Narrative Reports

Narrative reports are detailed, written accounts of what each student has learned in relation to the school's curriculum framework and the student's effort in class. The hope is that narrative reports will replace the shortcomings of letter grades because the latter tend to condense too much information into a single symbol. Narratives also allow teachers to include unique information about students' learning or something unique the teacher has done for that student—things that would not appear on a standardized form (Power & Chandler, 1998).

Advantages The concept of providing a rich description of a student's learning and educational development is laudable. When done well, these descriptions can mean much more to parents and students than the simple summaries that grades provide. This would be useful for describing elementary students' learning, especially if a state or school has defined a continuum of learning targets and performance standards over several grades, with benchmarks defined for each grade.

Limitations Narrative reports can be poorly or insensitively written, of course. Even teachers with good intentions find them difficult to write well. They may confuse or overwhelm parents, who may be asked to read 5 to 10 pages of narrative to understand what their children are learning. Using narrative reports should not be undertaken without considerable teacher development. A mean teacher can be just as mean in narrative writing as in letter grading: "Johnny thinks like a chicken!" Sensitivity and constructive comments are necessary. Lots of guided practice in writing nonthreatening and nonblaming comments is needed.

Modified Narrative Reports Because meaningful long narrative reports are very time-consuming for teachers to prepare, some schools have modified the reporting process. One way to do this is by combining the checklist or rating-scale procedure with short written comments about each student. Figure 14.6 shows one section of a primary school pupil narrative report. The full report is four pages and includes a few pages showing the school's educational developmental continuum (Egawa & Azwell, 1995).

You can see from the example that the indicators function in a way similar to checklists (even if you do not actually check them) and provide a kind of framework for interpreting the teacher's brief comments. The presence of the indicators reduces the need for a teacher to explain what curricular activities were used and evaluated for each student.

Along the same lines is the *standards-based report card*, developed in the Tucson Unified School System (Clarridge & Whitaker, 1997). Figure 14.7 shows an example. For each curriculum area, standards were written for Grades 1–2 and 3–6. Each standard was adopted from the state's standards and written to match the district's core curriculum. In that way, standards were linked to specific learning targets. If a student achieves a state's standard, the teacher gives the student a quality score of 4. Teachers also prepared verbal descriptions of levels 3, 2, and 1 for each standard to explain the meaning of lesser levels of achievement, much in the same way one would develop general scoring rubrics. All of these verbal descriptions were computerized using a database program.

Parent-Teacher Conferences

Conducting Conferences **Parent-teacher conferences** are one of the best ways to build strong connections with parents, to provide them with an understanding of their children's learning strengths and needs, and to help them be involved in their children's learning. However, you need to conduct them carefully and skillfully if they are to be successful. Figure 14.8 lists some of the things to do before, during, and after the conference to keep it on target. Additional suggestions are given in Shalaway's (1998) *Learning to Teach . . . Not Just for Beginners*.

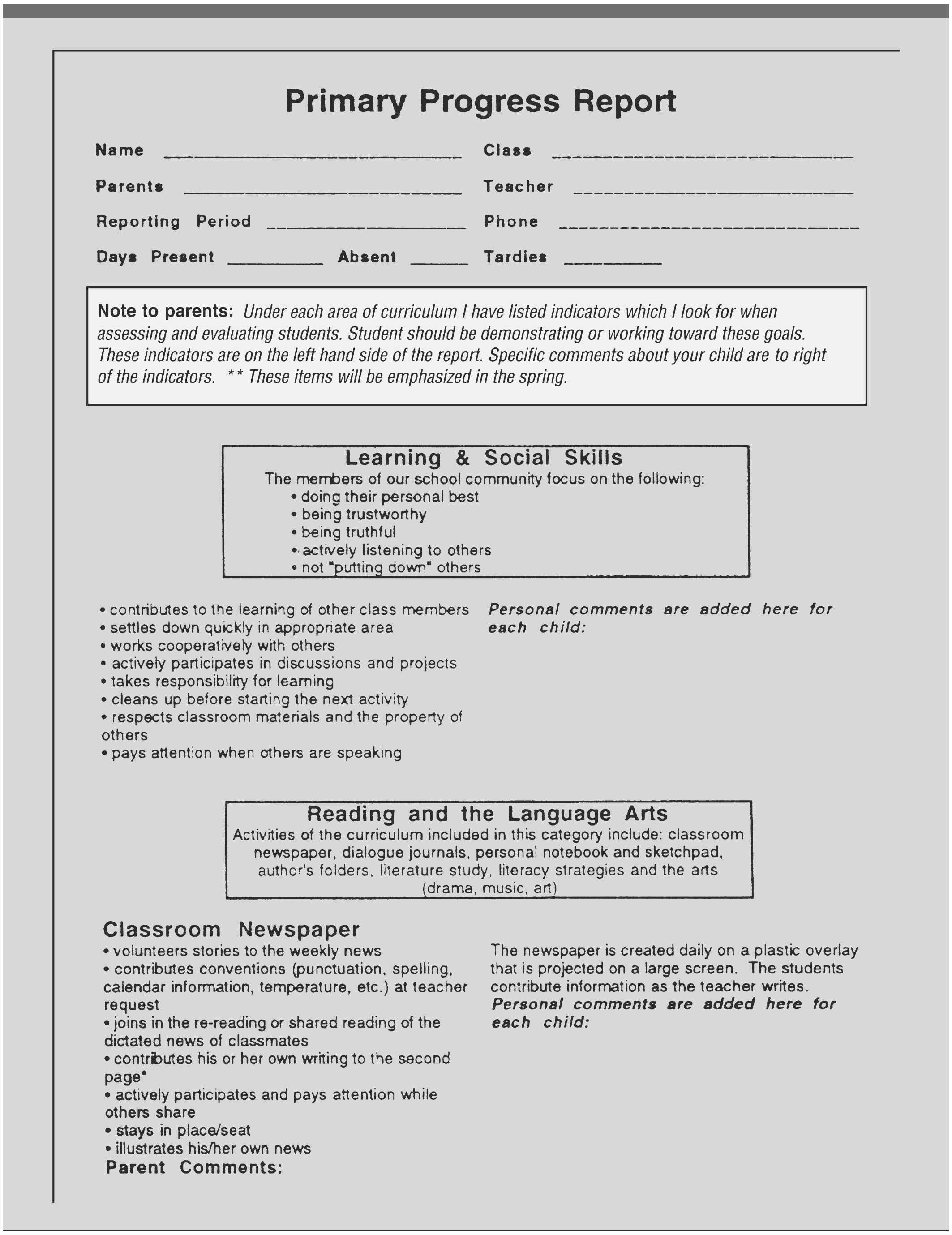
Limitations of Conferences Parent-teacher conferences have their drawbacks, however. They are time-consuming for the teacher, both in preparation time and in actual contact time. Schools frequently schedule 1 or 2 days for holding conferences during school hours; some schedule evening hours for the convenience of working parents. Sometimes schools neglect to give teachers time to plan and prepare for the conferences, assuming that teachers either need little or no planning time, or that they will do the necessary preparation after hours. In addition, parent-teacher conferences can be frustrating and produce anxiety for both teacher and parent, especially if the parties lack confidence in each other.

Attendance may also be a problem. Not all parents will come to conferences. Parents may be working, ill, embarrassed about their poor English or their poverty, unwilling to attend, or otherwise unable to come. Some parents are courteous and will notify you that they cannot attend, but you should not expect most parents to do this.

Finally, teachers and/or parents may have too much information, too many issues, or too many

FIGURE 14.6 **Example of a section of a primary-level narrative student progress report using indicators and teacher comments.**

Source: "Telling the Story: Narrative Reports," reprinted by permission of Kathy Egawa and Tara Azwell. In *Report on Report Cards: Alternatives to Consider* edited by Tara Azwell and Elizabeth Schmar. (Heinemann, a division of Reed Elsevier Inc., Portsmouth, NH, 1995.) Figure 9.1, p. 103.



Primary Progress Report

Name ______________________________ Class ______________________________

Parents ___________________________ Teacher ____________________________

Reporting Period __________________ Phone ______________________________

Days Present _________ Absent _____ Tardies _________

Note to parents: *Under each area of curriculum I have listed indicators which I look for when assessing and evaluating students. Student should be demonstrating or working toward these goals. These indicators are on the left hand side of the report. Specific comments about your child are to right of the indicators. ** These items will be emphasized in the spring.*

Learning & Social Skills

The members of our school community focus on the following:

- doing their personal best
- being trustworthy
- being truthful
- actively listening to others
- not "putting down" others

- contributes to the learning of other class members
- settles down quickly in appropriate area
- works cooperatively with others
- actively participates in discussions and projects
- takes responsibility for learning
- cleans up before starting the next activity
- respects classroom materials and the property of others
- pays attention when others are speaking

Personal comments are added here for each child:

Reading and the Language Arts

Activities of the curriculum included in this category include: classroom newspaper, dialogue journals, personal notebook and sketchpad, author's folders, literature study, literacy strategies and the arts (drama, music, art)

Classroom Newspaper

- volunteers stories to the weekly news
- contributes conventions (punctuation, spelling, calendar information, temperature, etc.) at teacher request
- joins in the re-reading or shared reading of the dictated news of classmates
- contributes his or her own writing to the second page*
- actively participates and pays attention while others share
- stays in place/seat
- illustrates his/her own news

Parent Comments:

The newspaper is created daily on a plastic overlay that is projected on a large screen. The students contribute information as the teacher writes.

Personal comments are added here for each child:

concerns to discuss in the brief time allotted to the conference. Often about 20 minutes is allotted for the conference. Also, some parents (and teachers) talk too much and use up more than their share of time. Scheduling another conference with the parents may be necessary.

Privacy Parent conferences should be private and between one teacher and the parent(s) of one student. The school principal should provide facilities to allow confidential discussions. Avoid holding a conference where other teachers, other students, or other parents can overhear what is being said. This protects the rights of all involved. It may be difficult to limit the conference to one teacher and the parent(s), especially in schools where students have different teachers for different subjects.

FIGURE 14.7 Example of a section of a computer-assisted narrative student progress report using standards-and rubric-based procedure along with teacher comments

Note: The report is for the first marking period (4 = the highest rating).

Source: Reprinted by permission from *Rolling the Elephant Over: How to Effect Large-Scale Change in the Reporting Process* by P. B. Clarridge and E. M. Whittaker. Copyright © 1997 by P. B. Clarridge and E. M. Whittaker. Published by Heinemann, a division of Reed Elsevier, Inc., Portsmouth, NH. All rights reserved.

		Semester			
		1st	**2nd**	**3r**	**4th**
Self-Directed Learner		**3**			
Student often sets achievable goals, considers risks, and makes some choices about what to do and in what order to do them, usually reviews progress, and often takes responsibility for own actions.	Comments:				
Collaborative Worker		**2**			
Student is developing the abililty to work in groups, has positive relationships with other students, and is learning to work toward group goals.	Comments:				
Problem Solver		**4**			
Student reasons, makes decisions, and solves complex problems in many situations, and uses these skills regularly, independently, and efficiently.	Comments:				

FIGURE 14.8 Suggestions for organizing and conducting a parent conference.

Source: Based on ideas from Brookhart (2009); Newman (1997–1998); Perl (1995); and Swiderek (1997).

SET PURPOSE

- Set goals for the conference.
- Decide what information you need to communicate with parents.
- Decide how, if at all, students will be involved, and what their role and tasks will be at the conference.

PLAN LOGISTICS

- If possible, send home report cards or other information about a week before, so parents have time to prepare questions and talk with their child.
- Schedule times and locations for each appointment. Include breaks for yourself at regular intervals.
- Keep to the schedule to respect everyone's time.
- Arrange for a waiting area where waiting parents cannot overhear your conference with other parents.
- Arrange a comfortable setting (chairs, tables, etc.) where you can converse easily.

COLLECT EVIDENCE

- Have grades, portfolios, student work samples, checklists, anecdotal records, etc., as appropriate, organized to share with parents. Work samples should illustrate the general level of student work and help parents understand their student's grades, current achievement level, and next steps.
- Involve students in the collection of evidence whenever possible.

INTERPRET EVIDENCE

- Prepare your main points ahead of time. Don't rely on spur-of-the moment thinking to convey important information about students. Clear oral communication requires just as much preparation as written comments do.
- Prepare questions you may have for parents about their child's work, interests, activities, etc.
- If you are well prepared, you can communicate clearly and remain confident.

FIGURE 14.8 (*Continued*)

COMMUNICATE

- Aim for clarity of expression; make your points clearly and briefly and support them with evidence.
- Listen carefully to what parents say. Respond to their concerns. Be open to learning more about the student than you know from the school setting.
- Use interpersonal skills: communicate genuine care for the student, develop rapport, and reflect parents' feelings.
- If the child is present, include him or her in the communication; if the child is not present, plan with parents how to share what went on so the child does not experience the conference as "people talking behind my back."
- Plan the next steps for the student jointly with parents.
- Do not allow antagonistic parents to derail communication. Your job is to understand the child's work and behavior as best you can, not to become the family's counselor or to become afraid or anxious. Listen and try to understand.

Multiple Marking Systems

A Report Card Example When a school uses more than one method to report students' progress, such as a report card with several kinds of marks or symbols, this is called a **multiple marking system**. A report card, especially for the elementary schools, usually uses a multiple marking system. Figure 14.9 shows an example of a **report card** employing a multiple marking system for Grades 4 through 6 in one school district.

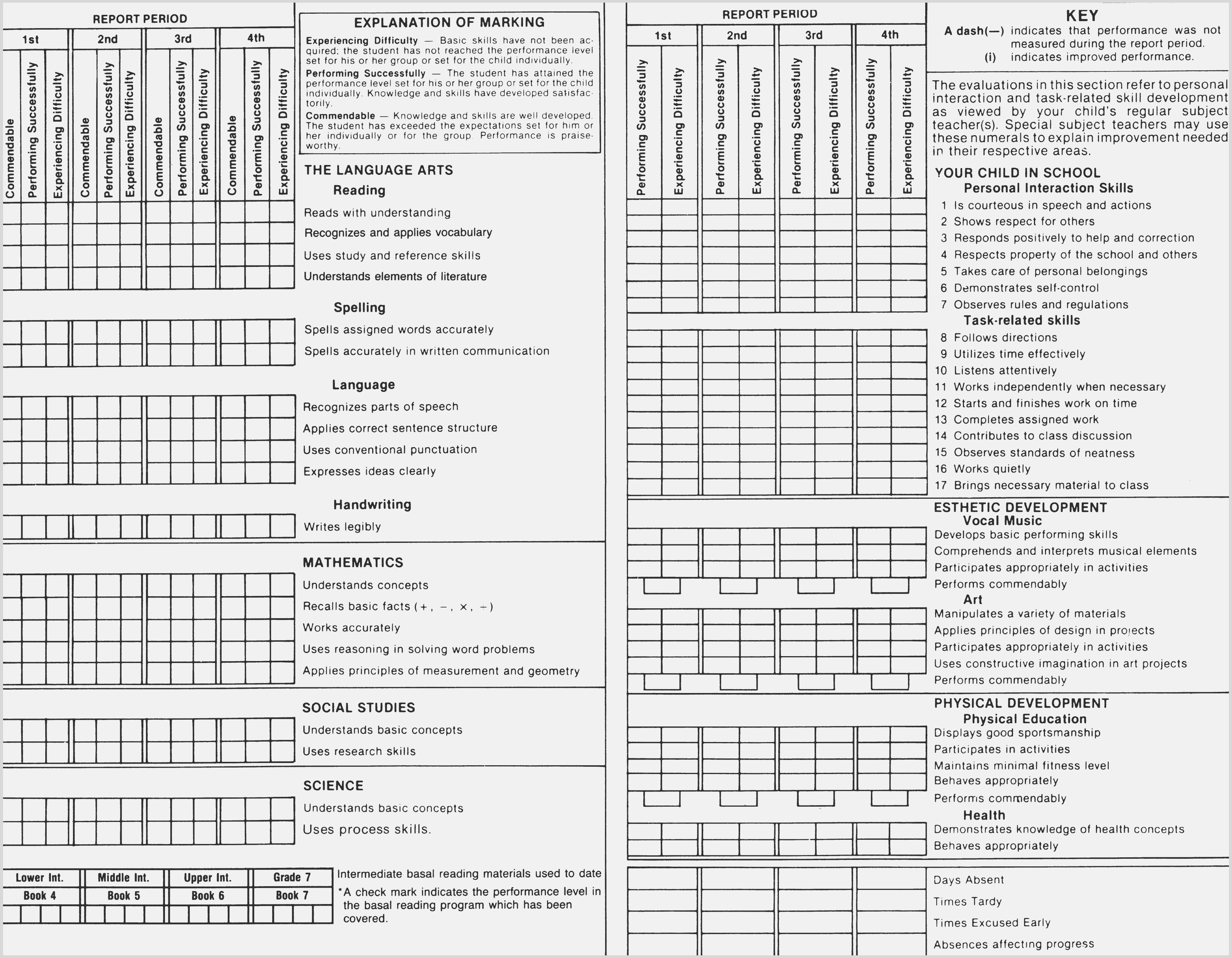
Reporting Achievement for Each Subject Notice the card has four marking periods, called "report periods," each approximately 9 weeks long. Words (*experiencing difficulty, performing successfully,* and *commendable*) define levels of accomplishment and serve as a rating scale for other areas. These are repeated as column headings under each marking period. For each marking period, the teacher uses a checkmark for reporting the student's achievement in each curriculum area. A dash (—) and an "I" also are used to communicate. In this report card, each curriculum area is divided into two to four subareas that contain the major learning targets of the curriculum in this school district. Reporting progress on each of them provides both parents and the following year's teacher with more specific information about what a student has achieved in the curriculum.

Reporting Noncognitive Achievement Progress in nonacademic areas is reported on the right side of the report card in Figure 14.9. Most schools rate citizenship, behavior, and so on separately from achievement, but this provision varies with the grade level. For kindergarten, primary, and upper elementary grades, most schools provide this separation; somewhat fewer schools do so at the middle school and senior high levels. Notice that in the example report card, the nonacademic areas are defined by specific, observable student performances. Thus, instead of asking teachers to rate general traits such as "personality" or "deportment," the school district asks the teacher to focus on specific student performances that can be observed and assessed.

Permanent Record A **permanent record** is the official record of a student's school performance. Not all information needs to appear on a student's permanent record card. Putting elementary students' letter grades in a permanent record is controversial. Many educators (and some professional associations) argue that reporting or recording grades at the elementary level is inappropriate. However, students and parents may become upset if, for the first time in middle school, a student receives a C (or lower) in a subject, when previously the student has received only "performing satisfactorily" checks on the elementary report card or a narrative report.

Some intermediate policy may help a student with this transition from the elementary school marking code to a new marking code at the junior high. A school may decide, for example, to have teachers prepare letter grades for fourth and fifth graders, but not to report them on report cards or on permanent record cards. Parents, however, are apprised of these grades. Thus, a "performing satisfactorily" can mean a C for some students and a

FIGURE 14.9 Example of a multiple marking system report card for Grades 4, 5, 6.



REPORT PERIOD												
1st			2nd			3rd			4th			
Commendable	Performing Successfully	Experiencing Difficulty	Commendable	Performing Successfully	Experiencing Difficulty	Commendable	Performing Successfully	Experiencing Difficulty	Commendable	Performing Successfully	Experiencing Difficulty	

EXPLANATION OF MARKING

Experiencing Difficulty — Basic skills have not been acquired; the student has not reached the performance level set for his or her group or set for the child individually.

Performing Successfully — The student has attained the performance level set for his or her group or set for the child individually. Knowledge and skills have developed satisfactorily.

Commendable — Knowledge and skills are well developed. The student has exceeded the expectations set for him or her individually or for the group. Performance is praiseworthy.

THE LANGUAGE ARTS

Reading

Reads with understanding

Recognizes and applies vocabulary

Uses study and reference skills

Understands elements of literature

Spelling

Spells assigned words accurately

Spells accurately in written communication

Language

Recognizes parts of speech

Applies correct sentence structure

Uses conventional punctuation

Expresses ideas clearly

Handwriting

Writes legibly

MATHEMATICS

Understands concepts

Recalls basic facts (+, −, ×, ÷)

Works accurately

Uses reasoning in solving word problems

Applies principles of measurement and geometry

SOCIAL STUDIES

Understands basic concepts

Uses research skills

SCIENCE

Understands basic concepts

Uses process skills.

Lower Int.	Middle Int.	Upper Int.	Grade 7
Book 4	Book 5	Book 6	Book 7

Intermediate basal reading materials used to date

*A check mark indicates the performance level in the basal reading program which has been covered.

REPORT PERIOD							
1st		2nd		3rd		4th	
Performing Successfully	Experiencing Difficulty	Performing Successfully	Experiencing Difficulty	Performing Successfully	Experiencing Difficulty	Performing Successfully	Experiencing Difficulty

KEY

A dash(—) indicates that performance was not measured during the report period.

(i) indicates improved performance.

The evaluations in this section refer to personal interaction and task-related skill development as viewed by your child's regular subject teacher(s). Special subject teachers may use these numerals to explain improvement needed in their respective areas.

YOUR CHILD IN SCHOOL

Personal Interaction Skills

1 Is courteous in speech and actions
2 Shows respect for others
3 Responds positively to help and correction
4 Respects property of the school and others
5 Takes care of personal belongings
6 Demonstrates self-control
7 Observes rules and regulations

Task-related skills

8 Follows directions
9 Utilizes time effectively
10 Listens attentively
11 Works independently when necessary
12 Starts and finishes work on time
13 Completes assigned work
14 Contributes to class discussion
15 Observes standards of neatness
16 Works quietly
17 Brings necessary material to class

ESTHETIC DEVELOPMENT

Vocal Music

Develops basic performing skills

Comprehends and interprets musical elements

Participates appropriately in activities

Performs commendably

Art

Manipulates a variety of materials

Applies principles of design in projects

Participates appropriately in activities

Uses constructive imagination in art projects

Performs commendably

PHYSICAL DEVELOPMENT

Physical Education

Displays good sportsmanship

Participates in activities

Maintains minimal fitness level

Behaves appropriately

Performs commendably

Health

Demonstrates knowledge of health concepts

Behaves appropriately

Days Absent

Times Tardy

Times Excused Early

Absences affecting progress

Source: Courtesy of the Mt. Lebanon, Pennsylvania, Public Schools.

B for others. At the end of the year, the letter-grades records are destroyed.

CHOOSING A GRADING MODEL

To make grades meaningful, adopt a grading framework for conceptualizing them and use that framework in a well-reasoned way consistent with your teaching approach. Grades must also be consistent with the reasons why you want to assign them and your school district's educational philosophy and policies. You may not have as much freedom as you might think in adopting a framework. Because grading students is serious business, to be professional you must choose and use a grading framework responsibly. You must be able to explain your grading framework to students, parents, and school officials. In this section we focus on making a decision about which grading framework to use.

Basic Teaching Approaches

Although there are many teaching methods and educational philosophies, most have a great deal in common. You may group teaching methods into two broad categories based on their major focus: learning target focus or performance of peers focus.

Focus on Defined Learning Targets This style of teaching focuses on having students attain high achievement by meeting high standards and achieving worthwhile learning targets. Teachers try to define standards and learning targets clearly and channel all efforts to achieving them. You may have heard of this approach under the name of one of its several variations such as standards-based (or standards-driven) instruction, performance-based instruction, or learning-objectives-based instruction. Teaching and instruction provide the conditions for students to meet the standards or learning targets.

Criterion-referenced grading frameworks are consistent with this teaching approach. Grades evaluate how well a student has achieved the specific learning targets or high standards. We are not focusing on where the students are in relation to where they began, nor are we focusing on how much further the students have yet to go before attaining the standards. Although these are worthwhile purposes, they are formative, not summative purposes. In criterion-referenced grading, we seek to sum up how much of the standards or learning targets students have achieved, or the quality level students have attained.

Focus on Performance of One's Peers Other educational approaches emphasize having students attain high achievement by outperforming their peers. The philosophy is that education should make one competitive; that the "cream rises to the top"; that all students are not capable of achieving high standards.

Norm-referenced grading frameworks are consistent with this teaching approach. Grades evaluate a student's achievement compared with other students' achievement: Where does a student rank in his or her class or in the school? Norm-referenced grading sums up achievement by communicating a student's success compared to other students.

There are actually three possible conceptual frameworks for grades. The three basic frameworks are the two mentioned above—*criterion-referencing* (absolute standards) and *norm-referencing* (relative standards)—and a third, *self-referencing* (growth or improvement standards). Figure 14.10 illustrates how grades can reflect the quality of a student's performance in relation to (a) quality levels describing achievement of learning targets or standards, (b) the performance of others in a specific group (such as classmates), or (c) the student's starting point or overall ability.

Criterion Referencing: Absolute Standards **Criterion-referenced grading** is also referred to as using **absolute standards grading**. You assign grades by comparing a student's performance to a defined set of standards to be achieved, targets to be learned, or knowledge to be acquired: Students who complete the tasks, achieve the standards completely, or learn the targets are given the better grades, regardless of how well other students perform or whether they have worked up to their potential. Thus, it is possible that you may give all students As and Bs if they all meet the absolute standards specified by the learning targets. Similarly, when you use this framework you must be prepared to assign all students Fs and Ds if none of them meet the standards set by the learning targets.

Criterion-referenced grading is most meaningful when you have a well-defined domain of performance for students to learn. The recent educational movement to set standards at the state level has put pressure on school districts to use these standards to set specific learning targets at the classroom level. The teachers in a school district often are left to align the specific learning targets with the standards. The aligned learning targets

FIGURE 14.10 Examples of definitions of grades under three different referencing frameworks.

Note: This figure is an adaptation of some of the ideas in Frisbie and Waltman (1992).

Absolute scale: task-referenced, criterion-referenced		Relative scale: group-referenced, norm-referenced	Growth scale: self-referenced, change scale
Grade	*Relative to the learning targets specified in the curriculum, the student has:*	*Relative to the other students in the class, the student is:*	*Relative to the ability and knowledge this student brought to the learning situation, the student:*
A	■ Excellent command of concepts, principles, strategies implied by the learning targets ■ High level of performance of the learning targets and skills ■ Excellent preparation for more advanced learning	■ Far above the class average	■ Made significant gains ■ Performed significantly above what the teacher expected
B	■ Solid, beyond the minimum, but not an excellent, command of the concepts, principles, strategies implied by the learning targets ■ Advanced level of performance of the learning targets and of most skills ■ Prepared well for more advanced learning	■ Above the class average	■ Made very good gains ■ Performed somewhat better than what the teacher expected
C	■ Minimum command of concepts, principles, strategies implied by the learning targets ■ Demonstrated minimum ability to perform the learning targets and to use basic skills ■ Deficiencies in a few prerequisites needed for later learning	■ At or very near the class average	■ Made good gains ■ Met the performance level the teacher expected
D	■ Not learned some of the *essential* concepts, principles, and strategies implied by the learning targets ■ Not demonstrated ability to perform some *very essential* learning targets and basic skills ■ Deficiencies in many, but not all, of the prerequisites needed for later learning	■ Below the class average	■ Made some good gains ■ Did not quite meet the level of performance the teacher expected
F	■ Not learned *most* of the basic concepts, principles, and strategies implied by the learning targets ■ Not learned most of the *very essential* learning targets and basic skills ■ Not acquired most of the prerequisites needed for later learning	■ Far below the class average	■ Made insignificant or no gains ■ Performed far below what the teacher expected

serve as the well-defined domain of performance students are expected to learn. Achievement of these learning targets becomes the basis for assigning grades. Arguments both for and against criterion-referencing, in general, center on whether it is of value to know exactly what the student has learned independently of the student's own capability and the learning of others.

Norm-Referencing: Relative Standards **Norm-referencing** is also called **grading with relative standards**. In this approach, you assign grades based on how a student's performance compared with others in the class: Students performing better than most classmates receive the higher grades. Advocates of group-referencing base their arguments on the necessity of competition in life, the value of knowing one's standing in relation to peers, and the idea that relative achievement is more important than absolute achievement. Arguments against norm-referenced grading center on the ill effects of competition, that the knowledge of standing in a peer group does not describe what a student has learned, and that ascertaining the absolute level of achievement is more important than ascertaining relative achievement.

With group-referenced grading, you must define the reference group against which you compare a student. Is the reference group the other students in this section of the course, in all sections taking the course this year, or all students taking the course during the past 5 years? Just as the criterion-referenced framework requires clearly defining learning targets for grades to be meaningful, so too does a group-referenced framework require clearly defining the reference group.

A grade based purely on a student's relative standing in a group does not convey to parents and school officials what the student is capable of doing relative to the curriculum's learning targets. Further, to act consistently within this framework, you should give good grades to the "top" students, even though they may not possess the level of competence specified by standards or the curriculum's learning targets. Similarly, you should give poor grades to the low-ranking students even though they may have met the minimum level of competence that the curriculum's learning targets specify.

Don't waffle and retrofit. You may start out wanting to grade using criterion-referencing and standards, but then discover that your students have done poorly. Being afraid to give poor grades, you may then waffle and start to "grade on the curve" (i.e., use norm-referencing). This retrofitting of a norm-referenced framework simply does not fit either approach and is not good educational practice. If the standards you set are grade appropriate and if students performed poorly, then you should determine why. Perhaps your assessment instruments were poorly designed (e.g., you may have used poor-quality testing materials that came with your curriculum). If so, then your assessments are invalid and no amount of norm-referencing can make them more valid. Perhaps your teaching was inadequate. Then reteaching is in order. Or perhaps the standards are simply not appropriate for the educational development of the students you teach. This is a matter that needs to be addressed by your principal or by the curriculum coordinator. In this case you need to adjust the standards, and then reteach: Grading on the curve in this instance distorts the real educational problem.

Self-Referencing: Growth Standards **Self-referencing** is also called **growth-based grading**. You assign grades by comparing students' performance with their own past performance or with your perceptions of their capability: Students performing at or above the level at which you believe them capable of performing receive the better grades, regardless of their absolute levels of attainment or their relative standing in the group. A student who came to the class with very little previous knowledge but who has made great strides may be given the same grade as a student who has learned more but who initially came to the class with a great deal more previous learning.

Arguments in favor of self-referenced grading center on the possibility of reducing competition among students and the concept that grades can be adjusted to motivate, to encourage, and to meet the students' needs. Arguments against the system center on the unreliable nature of teachers' judgments of capability, the need for parents and students to know standing relative to peers, the idea that this procedure tends to be applied mostly to lower ability students, and the possibility that this system may eventually lead to grading based solely on effort (Dunbar, Float, & Lyman, 1980). Additionally, students may not achieve the state's standards set for the grade.

From a statistical viewpoint, grading purely on growth or change may result in a negative correlation between the students' initial level of achievement and their growth: Students coming into class with the highest levels of achievement tend to have the smallest amount of measurable improvement or change, even though their final absolute levels of achievement remain the highest. This presents an irony: Students knowing most when they come into the course will tend to get the lowest grades because, even though in an absolute sense they may know more than most other students at the end of the course, they have shown a smaller amount of growth or change.

Your school district's grading policy and a grading culture are important factors in selecting a grading framework. Not every school district has a clearly written grading policy, but if your school district has a grading policy, you will be required to work within its guidelines. If it is a poor or inconsistent policy, you may wish to suggest ways to improve it. If you are a new teacher, your suggestions may not be taken seriously until the administration has confidence in your ability to teach. Press on with your reforms after you have taught for a year or two: Begin by working out your ideas with your most valued teaching colleagues. Don't ever give up on improving education for your students.

GRADING PRACTICES

This section focuses in more detail on using your assessment plan for summative grading. As you implement summative grading you must address at least seven issues so that your grades are valid:

1. Consider what types of student performance you should grade. We discuss three categories of student performances: those assessed, those reported, and those graded.
2. Consider how to make your marking scales consistent across all assignments throughout the marking period.
3. Decide the components making up the grade and their weighting in relation to the final grade.
4. Consider the standards or boundaries for each letter grade: How are they set and are they meaningful?
5. What about borderline cases? What do you do with students who are just at the border between two letter grades?
6. Be concerned with the issue of failures (Fs). What does failure mean?
7. Be concerned with the practice of assigning zero for a mark on one or more components going into a grade: What is the impact of this practice? When should a zero not be given?

Link Your Grading to Your Assessment Plan

In Chapter 6 we discussed how to craft an assessment plan. Your assessment plan describes what component assessments will make up the summative assessment for each instructional unit and for the marking period. In addition, you specify the weights the components will carry in the grade for each unit as well as the units' weights in calculating the final grade for the marking period. Figure 6.3 showed an example of this type of assessment plan. The assessment plan becomes critical to assigning grades. It enables you to integrate all the assessment components meaningfully into a valid grade and to explain your grading to students, parents, and school administrators.

What to Assess, Report, and Grade

Assessment Variables In Chapter 6 we discussed the types of student information you need when teaching, including sizing up the class, diagnosing students' needs, prerequisite student achievements, students' attitudes, students' work habits, students' study skills, and students' motivation and effort in school. The complete set of these characteristics for which you gather information are called **assessment variables** (sometimes called **evaluation variables**; Frisbie & Waltman, 1992). However, not all variables you assess need to be recorded and reported. Clearly, you will use some of the information to plan and guide your own teaching. This information is primarily formative. It should not make its way into a grade. A grade is a summative evaluation of a student's achievement.

Reporting Variables Your school district will expect you to report a subset of the assessment variables to parents and for official purposes. These are called **reporting variables** (Frisbie & Waltman, 1992). They often include the students' achievement in the subject, study skills, social behavior and interpersonal skills, motivation and study efforts in class, leadership skills, and aesthetic talents. This is illustrated by the multiple-marking system report card example shown earlier in this chapter.

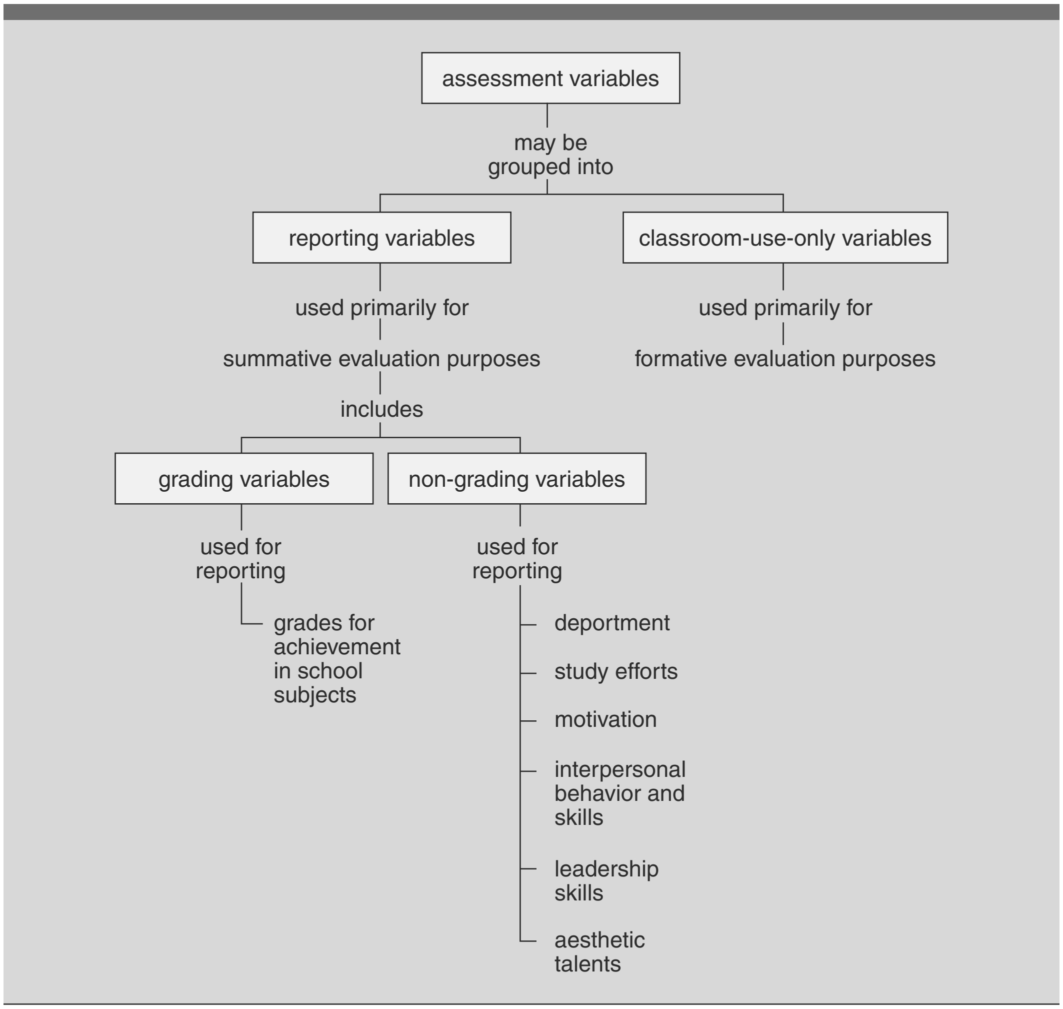
Grading Variables Reporting variables represent important school outcomes and therefore should be appropriately reported to parents and others. They should not be confused, however, with grades for course achievement. That is, from among all the reporting variables, there is a more limited subset on which you may base your grades. The variables in this limited subset are called **grading variables** (Frisbie & Waltman, 1992). You use the grading variables to describe a student's accomplishments in the subject. You assess these achievements by crafting more formal procedures such as performance tasks, portfolios, projects, tests, and quizzes. They are the most valid and reasonable bases for assigning grades.

Relationships Among Variables It is important that you be mindful of these variables as you assign grades. Figure 14.11 will help you understand the relationships among these variables and how they are used.

Eliminate Mixing If you mix grading variables with other variables, you create grades that have confusing and invalid meanings. For example, if you punish a student by lowering his or her grade for failing to turn in an assignment or for turning it in late, then you have confused the student's achievement with the student's behavior. Similarly, if you lower a science or social studies grade because of poor language usage or poor appearance, your grade is a less valid assessment of the student's achievement of the science or social studies curriculum learning targets.

This does not mean that language usage or turning in work on time is irrelevant to a student's school experience. Rather, the intention is clarity of meaning for grades so they become more valid indicators of achievement. Some schools, for example, use a "writing across the curriculum" approach. This means that social studies, history, mathematics, and science work is evaluated for both the subject-matter correctness and language usage.

FIGURE 14.11
Relationships among different types of assessment variables and grading variables.



Evaluations of the students' language usage are reported as part of the language grade, whereas evaluations of students' subject-matter achievement become part of the subject grade. Similarly, tardiness, failure to complete work, and other problems can be reported separately from achievement and may be used to explain a student's lack of school accomplishment.

Eliminate Formative Evaluation Components Not all achievement variables should be included as grading variables (Frisbie & Waltman, 1992). Many achievement variables are formative in nature. Homework, quizzes, and oral responses to classroom activities, for example, may serve mostly formative purposes—to help you decide whether individual students need more instruction, whether your lessons are going well, and to use as a basis for helping students to improve their performance. These formative assessments *should not* be included in the subject grade for the marking period. Not all out-of-class assignments are formative, of course. Some homework, most projects, and most research papers can be used for summative evaluation. The general rule, then, is to *include in the grade the assessments that you establish as useful for summative evaluation and exclude all assessments established primarily for formative evaluation*.

Craft Marking Scales to Be Consistent Across Different Assessments

Incompatibility of Scales Think ahead to make your assessment scales compatible across all the components that go into the summative grade. The assessment plan for the weather unit in Figure 6.2, for example, shows five components entering into the summative grade for the unit: homework, quizzes, independent investigation, map drawing, and the end-of-unit test. Suppose each of these is marked on a different scale as follows:

Component	*Scale*
Homework	0–10
Quizzes	0–5
Independent investigation	1–20
Map drawing	1–4
End-of-unit test	0–100

If you simply add students' marks from each of these components using these scales, you will have difficulty because they are incompatible. The map-drawing scale, for example, may be based on a rubric with four levels of quality whereas the end-of-unit test is based on a percentage scale from 0% to 100%. Such incompatibilities make a *simple sum of the marks* an invalid basis for a grade. You will need to mark each assessment in a way that makes scales compatible.

The planning stage is the time to prevent this situation. You may use one of several options, which we shall discuss later in this chapter. Solving this problem is not that complicated, but it is best solved up front. The following anecdote illustrates this point:

> In a school district I work with, eighth-grade teachers were faced with the task of combining percentage-correct scores from conventional language arts tests and writing performances scored on a 4-point rubric into five levels for report card grades (A, B, C, D, F). Several of the teachers did not have the quantitative reasoning background to understand why or how scale conversions could be made, and it had not occurred to any one of the several people who adopted the 4-point writing rubric that it would not be very helpful for assigning five levels of grades. This is a more complicated problem to solve after the fact than to solve at the design stage, when it would be appropriate to choose rubrics and construct decision rules. (Brookhart, 1999, p. 8)

Losing Precision The most reliable scores are those that are able to distinguish small differences in the quality of students' learning. A scale that allows you to demonstrate that Sally's command of a learning target is slightly better than Johnny's is more reliable than a scale that cannot tell the difference between their learning levels. To allow reliable detection of small differences between students, a score scale needs many gradations or "points." A scale that shows Sally at 89 and Johnny at 82 displays their relative learning better than a scale that shows them both receiving the same rating of B.

You lose precision when you transform scores from a fine-grained scale (e.g., percentage correct scale) to a coarse-grained one (e.g., letter grades). If a B were defined to be a score from 80 to 89, then both Sally and Johnny would receive the same grade, B. Because they both receive the same grade, their true difference cannot be distinguished with the letter-grade scale. By transforming the 89 and 82 both to a B, you have lost reliability.

Not all percentage scales are, in fact, fine-grained. For example, if you have five test questions, each worth 1 point, then the only possible percentages are 0, 20, 40, 60, 80, and 100. Thus, only six possible percentage values are used, not the 100 values you usually associate with a percentage scale. In this example, the percentage scale is just as coarse as the letter-grade scale. A test of 10 questions, each worth 1 point, is similarly not very fine-grained. Keep in mind that scales reporting fine differences among students must use many numerical values to be reliable assessments of the achievement differences among students. *If you use only a few of the many possible values of a scale, then you lose precision*.

Although you lose precision when you move from a fine-grained scale to a coarse-grained scale, you *do not gain precision by moving from a coarse-grained to a fine one*. If we have only the coarse scores initially, no transformation will make them more precise. Suppose, for example, you had the following writing scale: 4 = advanced, 3 = proficient, 2 = basic, 1 = below basic. Suppose your scoring rubric evaluated a student's writing as a 3 on this 4-point scale. You could transform the 3 to a percentage, with 3 out of 4 points becoming 75%. You have not gained any precision, however, in distinguishing among students because all students who received 3s now receive 75%. Unfortunately, from the precision standpoint, the scale has only 4 points after the transformation (25%, 50%, 75%, and 100%), the same as before the transformation. Only the labels have changed. In addition, because the 100% scale implies there are other possible percentages between those reported (especially between 75% and 100%), you have changed the meaning of the scale from advanced, proficient, basic, below basic (if those were the rubric levels) to an implied (from the percents) scale of A = 100%, C = 75%, F = 50%, and F = 25%. You can see that these so-called grades have a corrupted meaning—they are not aligned with the original meaning intended by the verbal labels of the writing scale.

Choices about scoring scales and precision can support or undermine the effects of even well-designed assessment tasks. It is important, therefore, to design your scoring scales as carefully as you design your assessment instruments.

Weighting the Components

You decide how much weight to assign to the components of a grade—home assignments, tests, quizzes, term papers, and other elements—after you decide their importance to the description of students' achievement of the learning targets. Begin by making a list of all the components you want to use for evaluating achievement of the grade. Next, decide how these components relate to the learning objectives and determine how important each is (and thus how heavily each will weigh) in relation to the overall summative grade.

Consider at least six factors when deciding how much to weight each component:

1. Components that assess more of the important learning targets and content should be weighted more heavily than those that focus only on one or a few targets.
2. Components that focus on what you spent the most time teaching the students should receive the most weight in determining the grade.
3. Components that require students to integrate and apply their learning should receive more weight than those that require students simply to recall what was taught.
4. When two components assess some of the same learning targets, each should be given less weight individually than other components that assess an equal number of unique learning targets (i.e., nonoverlapping components; Frisbie & Waltman, 1992).
5. If you know that one of the components you want to count toward the grade has some degree of unfairness to certain groups of students, you should be extremely cautious in using it for grading. If you decide that on the whole it is still appropriate to use it, you should weight it less, especially for students for whom it is less fair. For example, you may find that a timed, written test does not adequately assess students with certain disabilities. In such cases, it would be appropriate to weight this procedure less for these students and to give other, more appropriate procedures more weight in determining their grades.
6. Components that are less reliable and less objective should be weighted less heavily than those that are more reliable and objective. However, this is not to say that you should avoid using less objectively scored assessments such as essays and portfolios for assigning grades. Rather, you should use scoring rubrics for marking them so the marks are more reliable.

Standards or Boundaries Between Grades

An important practical consideration is how to establish boundaries between the grades. What constitutes an A, B, and so on? The answer will depend on the reference framework you are using and your school district's policy. The procedure for setting norm-referenced grading boundaries is quite different from the procedure for setting criterion-referenced grading boundaries.

Your grade boundaries must have the same meaning across all assessments that will make up the grade. This doesn't mean that you need to use the same number of marks (points) for each assessment. It does mean, however, that an A on one assessment should be of approximately the same standard of quality across all assessments. For example, if each assessment is marked according to the percentage correct, then the same percentage range (e.g., 90%–100%) should be used for an A across all assessments, and the quality of work represented by these percentages should be comparable across assessments.

Borderline Cases

You will always have **borderline cases**—students whose composite marks are very near or right on the boundary between two grades. Should you consider adjustments? How close to the grade boundary does a student have to be before you adjust a letter grade upward or downward? Many teachers are comfortable reviewing students' work and raising grades for those who are just under the borderline, but do not consider lowering the grades of those just above the borderline (Brookhart, 1993). Nevertheless, lowering borderline grades is just as valid as raising them when additional achievement evidence justifies it.

As you learned in Chapter 4, assessment results contain errors of measurement, so students whose scores are on or near the border are likely to have true scores that are *different* from their observed scores. This argues against being hard-nosed and telling a student that he or she missed the next higher grade by 1 or 2 points. You can think of scores near the grade boundary as in an "uncertainty band" much like the one discussed in Chapter 4. You should use additional achievement information about the student to help you decide whether the student's true score is above or below the boundary. Using additional *achievement* information to help make boundary decisions is more valid than using information about how much effort a student put forth in studying (Brookhart, 1999). If you are still in doubt, it is better pedagogy to give the next higher grade than to give the lower grade.

The Meaning of Failure

As Frisbie and Waltman (1992) point out, the grade F carries a lot of emotion with it because there are usually negative consequences for students who receive it. What should an F mean? Your answer should be consistent with your grading framework. The least confusing way to assign a failing grade (F) is to set reasonable minimum standards regarding performance on the curriculum learning targets. Students who *consistently* perform below these minimum performance standards receive an F.

Consider two students. Darnell does not turn in an important assignment, even though he knew the deadline and you made several announcements in class. You decide to give Darnell a zero. James, on the other hand, turns in the assignment on time, but the work is so poor you must give it a 55, which is in the F range. Both James and Darnell receive Fs. The question is, do these Fs mean the same thing? If not, how meaningful (i.e., valid) is using an F? We will address this issue momentarily, but first another example.

Many scoring rubrics for performance assessments and constructed-response items use a scoring scale from 1 to 4. Our concern is with the meaning of the lowest category, 1. Usually, the rubric describes 1 as very poor quality work, amounting to failure. However, the 1 is often assigned to students who did nothing, failed to turn in the work, or wrote gibberish. Does a score of 1 mean the same for every student? If not, how valid is it? This issue is not necessarily resolved by using a scoring rubric that goes from 0 to 4. If the zero is given to students who failed to turn in an assignment, as well as to those who turned it in but wrote poorly, the zero has two different meanings. This type of confusion lowers the validity of the marks.

One way to frame this issue is to consider two categories of student performance (Brookhart, 1999): (1) doing work that is of very poor quality, that is, failing work and (2) not doing the work at all, that is, failing to try. The first category describes the student's achievement: the student's status compared to the standards or learning targets. The meaning of failure marks or grades for such students is reasonably clear.

The second category reflects a student's motivation (and perhaps attitudes and personality characteristics, such as lack of self-confidence, test anxiety, rebelliousness, etc.). Darnell's failure to turn in an assignment might be a signal to you that he has not understood what you taught. Darnell may have failed to do the assignment because he didn't know how. This calls for working with Darnell and his parent(s) to see that he receives the help he needs. Darnell may be insecure and afraid to admit his failure to learn: Not every failure to try is malicious. Sometimes children who have an emotional crisis at home actually do the work in school but do not turn it in because they have given up being successful students. In "failure-to-try" cases, giving a failing grade (or lowering a grade) is always invalid because the resulting grade does not accurately describe achievement. This does not mean that you should avoid reporting failure to try; it does mean that describing these two types of student responses with the same mark (0) or with the same grade (F) is not valid.

A closely related question is, "Should I lower a student's grade when the assignment is turned in *late*?" Some teachers, for example, mark assignments that are turned in late, but deduct points from the mark or otherwise limit the highest mark possible for this assignment. Again, such a practice lowers the validity of the marks and the resulting grades because it mixes up their meaning: Do not use the same grade to describe for some students only achievement, but for other students a mixture of achievement, attitudes, and personality evaluations.

Abhorrent grading practices like these are practiced because teachers face difficult teaching conditions. They seek to use grades (and student evaluations, in general) to control students' behavior. As we discussed in Chapter 5, it is poor practice to threaten, punish, or manipulate students by lowering achievement grades for behavior that is unrelated to achievement. The issue of what to do with missing and late assignments is a real one with which you and your colleagues must struggle, but it is not a measurement problem per se. It is a result of the conditions of teaching, school policies, and assumptions people make about the way one should educate (Brookhart, 1999).

A school district's policy needs to address how to handle students who do not turn in assignments or who turn them in late. A culture for punctuality and completing assigned work on time needs to be developed. A policy needs to be legal, fair, and valid, and it needs to meet criteria for sound educational philosophy. Punishing, threatening, or manipulating students should be eliminated from any policy.

Strictly from a measurement point of view, assigning an "incomplete" when assignments are not turned in seems reasonable. Students who complete the work beyond the deadline may be given full credit from a measurement perspective. The report card could contain a notation that some of the work on which the grade is based was completed after the due date. Repeated notations of this type describe tardiness but do not detract from describing achievement.

These measurement "solutions," however, do not address all the concerns of teachers. You can raise questions such as: Is it fair to students who habitually complete their work on time to allow other students not to complete theirs on time? Are there circumstances under which late work is allowed (without penalty or commentary) or appropriate (e.g., illness, personal tragedy)? Will a flexible policy on when to turn in work result in classroom chaos? Is the assignment of an invalid grade or a grade with low validity more ethical than addressing the issues of why students do not behave properly or do not turn in assignments? You may tackle these issues with your instructor and teaching colleagues.

The Deadly Zero

Do you recognize how much a zero can affect a composite score? Suppose Ashley is a good student, capable of B work. What happens to her average marks if she fails to turn in one assignment and you give her a zero for it? Are you surprised to learn that her average grade could drop from a B to a D?

The impact of a zero, of course, depends on the component marks a student receives, how many marks enter into the composite grade, the weights assigned to the component, and the mark the student would have received had she turned in the assignment. Figure 14.12 may help you understand the impact of zero on a student's grade.

In this example, there are five assignments. To keep things simple, let us assume they are equally weighted. As a point of reference, suppose Ashley's "true performance," what she would have received had she completed all her assignments, is shown in Panel A. Ashley is a B student.

FIGURE 14.12
Hypothetical example of the impact of substituting zero or 59 for one assignment a student did not turn in.

Note: Substituted values are shown in parentheses. (Assume A = 90–100, B = 80–89, C = 70–79, D = 60–69, F = 0–59.)

	1	2	3	4	5	Avg	Grd
A. True Performance	80	70	85	75	90	80	B
B. Strategy 1—Substitute zero for the missing assessment							
Case 1	(0)	70	85	75	90	64	D
Case 2	80	(0)	85	75	90	66	D
Case 3	80	70	(0)	75	90	63	D
Case 4	80	70	85	(0)	90	65	D
Case 5	80	70	85	75	(0)	62	D
C. Strategy 2—Substitute the highest possible failing mark (i.e., 59) for the missing assessment							
Case 1	(59)	70	85	75	90	76	C
Case 2	80	(59)	85	75	90	76	C
Case 3	80	70	(59)	75	90	75	C
Case 4	80	70	85	(59)	90	75	C
Case 5	80	70	85	75	(59)	74	C
D. Strategy 3—Base the grade on only those assignments that were turned in							
Case 1	—	70	85	75	90	80	B
Case 2	80	—	85	75	90	83	B
Case 3	80	70	—	75	90	79	C
Case 4	80	70	85	—	90	81	B
Case 5	80	70	85	75	—	76	C
E. Strategy 4—Substitute zero for the missing assignment, and use the median to calculate the grade							
Case 1 (0)	(0)	70	85	75	90	80	B
Case 2	80	(0)	85	75	90	83	B
Case 3	80	70	(0)	75	90	78	C
Case 4	80	70	85	(0)	90	82	B
Case 5	80	70	85	75	(0)	78	C

Panel B shows what will happen to Ashley if she fails to turn in one assignment and if you were to give her zero for that assignment. The impact on her grades is dramatic: One missing assignment results in her dropping two whole grades, from a B to a D. This happens no matter which assignment she fails to turn in.

Using a zero means that you have given Ashley the lowest possible failing mark as a substitute for her missing assignment. Instead, you could give her the highest possible failing mark. In this example, the F range is from 0 to 59, so 59 is the highest possible failing mark. Panel C shows what happens to Ashley's grade if you follow this strategy. Ashley goes from a B average (Panel A) to a C average (Panel C). Still, one missing assignment has resulted in her average dropping one whole grade.

Panel D shows what happens when you simply ignore the missing assignment, basing your grade on the remaining four. As shown in Panel D, the impact on her grade depends on which assignment she failed to turn in. If she failed to turn in one of the two on which she could have scored the highest (Case 3, where Assignment 3 was 85, or Case 5, where Assignment 5 was 90), her grade would drop a whole grade; in the other cases it would remain at B. Other strategies (not shown) could be used, such as using 50 instead of 59 or substituting the average of the four completed assignments for the missing assignment. In Ashley's case these other approaches give the same letter-grade results as shown in Panel D.

Panel E shows what happens when you do give Ashley the zero for missing assignment, but instead of using the mean of the grades to calculate the average, you use the median. Appendix I shows how to calculate the median, which is a good measure of central tendency to use for distributions that include extreme scores, like the zeros here.

From the measurement perspective, Strategy 3 (basing the grade only on assignments turned in) would be the best of the three when (a) assignments are of approximately equal difficulty for the students, (b) assignments are weighted equally (or are worth the same number of points), and (c) there are several assignments and only one or two are missing. This recommendation does not consider other factors, such as whether (a) the "missing assignment" is the most important one to complete (e.g., a project or a final examination), (b) a student fails to turn in an assignment because of illness or personal tragedy, (c) a student fails to complete the assignment because she didn't understand how to do the work, and (d) a student has made a habit of not turning in work on time. As we stated previously, these are not measurement issues per se but matters of educational practice, classroom management, and school policy.

From a practical perspective, we recommend Strategy 4 if it is not possible for you to ignore a missing assignment. Sometimes grading policies or timelines require a set of work to be considered at a certain time. If repeated efforts to help students turn in assigned work have not yielded results, using the median method of calculation allows you to take the zero into account but not give it undue weight. However, we recommend that you use the same calculation method for all students in the same class, so if you use the median for one student in a class you should use it for all of them.

TECHNIQUES FOR SETTING GRADE BOUNDARIES AND COMBINING SCORES

Assigning Norm-Referenced Letter Grades

We have recommended criterion-referenced grading (see next section) to match with standards-based instruction. We include this section on norm-referenced grading for the sake of completeness, because there may be occasions when norm-referenced grading methods are required. Your instructor may ask you to skip this section, or to postpone studying this section until you have studied Chapter 16.

Several methods of assigning grades use relative or norm-referenced standards. One method, called **grading on the curve,** uses the rank order of students' marks: Students' marks are ordered from highest to lowest, and grades (A, B, C, etc.) are assigned on the basis of this ranking. A second, called the **standard deviation method,** uses the standard deviation (see Appendix I) as a unit: A teacher computes the standard deviation of the scores and uses this number to mark off segments on the number line that define the boundaries for grade assignment. The two methods do not necessarily give the same results. We explain how to use the methods next.

Grading a Single Test or Assessment There are several methods of grading a single test or assignment in a norm-referenced manner. Most of these were devised before criterion-referenced assessment became the method of choice and are now somewhat dated. We present only one such method here, known as grading on the curve. To use the grading on the curve method to assign letter grades, you decide on the percentage of As, Bs, Cs, and so on to award. For example, you may decide as follows:

Example

Example of one possible set of percentages to use for grading on the curve

Top 20% of the students get A

Next 30% get B

Next 30% get C

Next 15% get D

Lowest 5% get F

There are no rules on how you would select the percentages to use. They are chosen arbitrarily based on your experience as to what is realistic in your school for a distribution of letter grades. This approach does not require using a normal or bell curve.

Another way to set the percentages is to divide the range of a normal or bell curve (see Chapter 16) into five equal-length intervals. The example below shows the resulting percentages of students receiving each grade.

Example

Example of one possible set of percentages to use for grading using the normal curve with five equal intervals

Top 3.6% of the students get A

Next 23.8% get B

Next 45.2% get C

Next 23.8% get D

Lowest 3.6% get F

This set of percentages assumes that the true achievement in the group of students in your class is normally distributed, an assumption which, in the authors' view, is hard for you or any teacher to justify. Notice that (a) the width of the interval that determines the percentages is completely arbitrary, (b) the assessment scores must be valid measures of the desired achievement, and (c) there is no reference to the learning targets, skills, or competence the letter grades represent (except that higher-ranked students have more competence than lower-ranked students). If you do decide to grade on a curve, then you must provide a convincing and educationally sound argument to justify the validity of the particular percentages that you use; otherwise, your grades are likely to be unsound.

Grading a Composite of Several Scores Usually a report card grade reflects a student's performance on several assessments such as assignments, quizzes, reports, and perhaps an examination. Here we discuss how to combine the scores from several grading components into a single (composite) mark in a manner consistent with the norm-referenced grading framework. We discussed previously the factors you need to consider when assigning weights to each component of the grade. There is no agreement as to exactly what weights are proper for each grading component.

Weighting Guidelines When norm-referenced grading is adopted, the component that contributes the most to the final rankings of the students in the group carries the most weight. This principle is likely to be violated if you simply multiply the component scores by some arbitrary weights and then add the weighted scores to form a composite. The reason is that the rank of a composite score is influenced by the standard deviations of the components making up the composite (and by the intercorrelations among components). To illustrate this, consider the next example:

Example

Hypothetical example showing how a grading component can work in the opposite way the teacher intends when norm-referenced grading is used

Suppose that the final grades are based on the sum of the marks from one exam and one project. Suppose further that the project is intended to weigh twice as much as the exam. In an attempt to accomplish this, the teacher decides to give twice as many points to the project as to the exam: 100 points for the project and 50 points for the exam. Remember that in a norm-referenced grading framework, those who rank highest should receive the highest grades. Here are the marks and ranks of five students.

	Exam (50 points)		Project (100 points)		Total	
Student	Marks	Ranks	Marks	Ranks	Marks	Ranks
Anthony	44	1	77	5	120	1
Ashley	33	2	78	4	111	2
Billy	26	3	79	3	105	3
Chad	22	4	80	2	102	4
Vanessa	15	5	81	1	96	5

Notice that the project ranks students exactly opposite from the exam. The final order is exactly the same as the exam, however, even though the teacher weighted the project more. This is because the ranking of the students on the total marks depends on the spread of scores rather than on the teacher's intended weighting. The spread of scores is measured by the standard deviation (see Appendix I). The project scores are close to each other, so their standard deviation is small, whereas the exam scores are quite different from each other, so their standard deviation is large. Because of the exam score's larger standard deviation, the students' *exam ranking* dominates their final total ranking in spite of the teacher's intention to make the project the dominant component. In general, when using norm-referenced grading, the larger the standard deviation of one component's scores, the more that component influences the final ranking of students when a composite is formed.

Using SS-Score Method The **SS-score method** preserves the influence (weights) you want the components to have, by first adjusting the values of the components' standard deviations. After adjusting, you may then apply the weights you desire. There are three steps: First, change all of the scores on each component into *SS*-scores (*SS* means linear standard score; see Chapter 16). This makes all of the standard deviations equal. Second, multiply the components' *SS*-scores by the weights you want. Finally, add these products to form the composite mark for each student. The following formula summarizes these steps.

$$\text{weighted composite score} = \Sigma(\text{weight} \times SS) \quad \text{[Eq. 14.1]}$$

where

$$\text{weight} = \text{weight you want the component to have}$$

$$SS = \text{the linear } SS\text{−score for the component}$$

$$= [\,10(X-M)/SD\,] + 50$$

The procedure is illustrated in Figure 14.13.

Note that these weighted composite scores are not themselves *SS*-scores. However, the weighted composite scores do provide a way to rank students so that the component weightings you specify will have the desired influence on the students' final standings.

To appreciate the influence of the *SS*-score method on the students' ranking in the final weighted composite, recall our earlier example in which the teacher's attempt to make the project dominate the ranking of the students failed. In the example below, we'll apply the *SS*-score method to those same marks.

Assigning Criterion-Referenced Letter Grades

There are several methods for grading using the criterion-referencing grading framework. In this book we shall discuss only three. One method is known as the **fixed-percentage method**: The scores on each component entering into the composite are first converted to percentage correct (or percent of total points); then the percentages are translated to grades. For each component, you must use the same percentage to define the letter-grade boundaries.

A second method is called the **total points method**: Each component included in the final composite grade is assigned a maximum point value (e.g., quizzes may count 10 points, exams may count a maximum of 50 points each, and projects may count a maximum of 40 points each); the letter grades are assigned based on the number of total points a student accumulated over the marking period.

A third method is the **quality-level method** or the **rubric method**. It is sometimes called the **content-based method** (Frisbie & Waltman, 1992). In this method, you describe the quality level of performance a student must demonstrate for each letter grade—what types of performance will constitute an A, B, C, and so on. (An example of these definitions of quality is shown in Figure 14.10 in column one.) Given these definitions, you evaluate the student's work on each component, decide the quality level of work, and then assign the corresponding grade. This method is very similar to using scoring rubrics for performance tasks (see Chapter 12). When you develop rubrics for a component, you must be sure the number of quality levels corresponds to the number of letter-grade levels.

Grading a Single Test or Assessment: Fixed-Percentage Method Teachers frequently use percentages as bases for marking and grading papers. The relationship between percentage correct and letter grade is arbitrary. In some schools,

Example

Hypothetical example showing using the SS-score method can help the teacher weigh the assignments as intended when norm-referenced grading is used.

	Exam (Weight = 1)			*Project (Weight = 2)*			*Total*	
Student	*Marks*	*SS*	*Ranks*	*Marks*	*SS*	*Ranks*	*Composite*	*Ranks*
Anthony	44	66	1	77	36	5	138	5
Ashley	33	55	2	78	43	4	141	4
Billy	26	48	3	79	50	3	148	3
Chad	22	44	4	80	57	2	158	2
Vanessa	15	37	5	81	64	1	165	1

Compare the final rankings in this example with the final rankings in the earlier example. Now the project dominates the rankings based on the composite instead of the exam. This result is what the teacher initially intended. Transforming the marks to *SS*-scores first made the exam marks' and project marks' standard deviations equal. Then when the weight of 2 was applied, the composite better matched the teacher's intent. You can use the standard deviation formula in Appendix I to work through this example yourself, if you wish.

FIGURE 14.13 **Example of calculating composite marks using the *SS*-scores method.**

	Components entering into the grade			
	Quizzes	**Homework**	**Term paper**	**Exam**
Mean (*M*)	70	85	75	65
Standard deviation (*SD*)	5	8	15	20
Teacher's weight	20%	10%	20%	50%
Calculation for *SS*[a]	*SS* = 10(Mark – 70)/5 + 50	*SS* = 10(Mark – 85)/8 + 50	*SS* = 10(Mark – 75)/15 + 50	*SS* = 10(Mark – 65)/20 + 50

[a] *SS*-scores are calculated by subtracting the component mean from a student's raw score, dividing the difference by the standard deviation, multiplying by 10, and adding 50 to the product. The results for each student are shown below. For example:

The quizzes *SS*-score for Bob = 10(87 – 70)/5 + 50 = 84
The quizzes *SS*-score for Chad = 10(85 – 70)/5 + 50 = 80
The quizzes *SS*-score for Susan = 10(75 – 70)/5 + 50 = 60
The quizzes *SS*-score for Theresa = 10(70 – 70)/5 + 50 = 50

	Raw scores on components				*SS*-scores on components				
Students	**Quizzes**	**Homework**	**Term paper**	**Exam**	**Quizzes**	**Homework**	**Term paper**	**Exam**	**Weighted composite**
Bob	87	85	70	80	84	50	47	58	60
Chad	85	80	80	70	80	44	53	53	58
Susan	75	82	85	60	60	46	57	48	51
Theresa	70	78	75	65	50	41	50	50	49

Composite scores are calculated by multiplying the component *SS*-score by the corresponding teacher's component weights and summing the products. For example:

Composite score for Bob = .2(84) + .1(50) + .2(47) + .5(58) = 60
Composite score for Chad = .2(80) + .1(44) + .2(53) + .5(53) = 58
Composite score for Susan = .2(60) + .1(46) + .2(57) + .5(48) = 51
Composite score for Theresa = .2(50) + .1(41) + .2(50) + .5(50) = 49

Source: From *Essentials of Educational Measurement* (3rd ed., pp. 248–251), by R. L. Ebel, 1979, Englewood Cliffs, NJ: Prentice Hall. Copyright 1979. Adapted by permission of the copyright holder.

80% is an A; in others, 85% is an A. In still others, 90% is an A. Some school boards have a policy on this matter. The following is an example of one such set of percentages that defines letter grades:

Example

Example of one possible set of percentages for grading using the fixed-percentage method

90–100% = A

80–89% = B

70–79% = C

60–69% = D

0–59% = F

Note that a percentage begs the question, percentage of what? Often, the only answer that you can defend is that the score represents the percentage of the maximum points on the test or the assignment. This answer ignores the broader concern: The test should be a representative sample from a well-defined domain of performance implied by the curriculum learning targets. If you have not defined this domain and have not built the assessment to sample the domain representatively, then you cannot use the percentage grade to estimate the student's status accurately on that broader domain. Such tests (or assessments), and consequently such percentages, cannot be considered criterion-referenced.

The percentage that defines each grade should take into account a teacher's experience with the kinds of students being taught and the difficulty of tests the teacher develops. Thus, norm-referenced information helps establish a criterion-referenced grading system.

One limitation of this fixed-percentage method stems from the fact that every assessment you create has a different level of difficulty, which you may not know in advance. This method, however, uses the same fixed percentages for A, B, and so on for every component. Thus, if you create a test that is too difficult for your class, you may end up giving too many low grades based on the percentages you fixed in advance. This will be frustrating for students and may put you into a position where you have to change the grading system.

A second limitation is that this method encourages you to focus more strongly on the difficulty level of the assessment than on the learning targets it should assess. For example, if you fix the percentages, you will be looking for ways to make the assessment easy enough or difficult enough so that you get a reasonable distribution of letter grades for your class. This seems to go against the principles of absolute or criterion-referenced grading.

Grading a Single Test or Assessment: Total Points Method To use this method, you must decide in advance all the components that will enter into the end-of-a-marking-period grade. Then, also in advance, you decide the maximum number of points for each component. Your assessment plan should do this. The maximum number of points each component is worth mirrors the weight you assign to each component. If you want the unit test(s) to count more toward the grade, for example, you would assign the unit tests more of the total points. Finally, you sum all the maximum points for components and use that maximum possible total to set letter-grade boundaries. Notice that, unlike for the fixed percentage method, you do not assign letter grades for each component, but only for the total summed over all components.

As an example, suppose you used the same four components that were used in one of our earlier examples: quizzes, homework, a term paper, and an exam.

Example

Example of one possible set of points to use with the total points method of grading

Component	Maximum points	"Weighting" expressed as a percentage
Quizzes	40	20%
Homework	20	10%
Term paper	40	20%
Exam	100	50%
Total Points	200	100%

Having decided on the components and their maximum point values, you then set the boundaries for assigning letter grades to the total points that students accumulate in the marking period. For example,

Total point grade boundaries	Grade
180–200	A
160–179	B
140–159	C
120–139	D
0–119	F

Notice that these total point grade *boundaries* correspond to percentages of 90%, 80%, 70%, and 60% of the 200 total points for A, B, C, and D, respectively. (For example, for an A, 180 ÷ 200 = 0.90 or 90%.) You may use other percentages to define the letter-grade boundaries. Adjust the total point boundaries accordingly.

One limitation of the total points method is that it makes it too easy for you to give "extra credit" assignments to boost the total points of low-scoring students. Extra credit assignments tend to distort the meaning of the grades, especially when these assignments do not properly assess the same learning targets as the original set of components. For example, if a student does poorly on the term paper, you may be tempted to have the student read and summarize a current events magazine article to boost the student's score instead of writing another term paper. The meaning of the total points for this student would be distorted relative to other students. As a result, your grades are less valid.

Another limitation of this method is that by defining the maximum number of points before creating the assessments, you may be faced with an unacceptable choice when you do create an assessment tool. Consider this situation:

> Suppose I need a 50-point test to fit my [total points] grading scheme, but find that I need 32 multiple-choice items to sample the content domain thoroughly. I find this unsatisfactory (or inconvenient) because 32 does not divide into 50 very nicely. (It's 1.56!) To make life simpler, I could drop 7 items and use a 25-item test with 2 points per item. If I did that, my points total would be in fine shape, but my test would be an incomplete measure of the important unit objectives. The fact that I had to commit to 50 points prematurely dealt a serious blow to obtaining meaningful assessment results. (Frisbie & Waltman, 1992, p. 41)

Grading a Single Test or Assessment: Quality-Level Method When you grade an individual assignment with a rubric or grading scale, you make a judgment based on the quality level of the work, overall or according to several criteria. In fact, as you saw in Chapter 12, performance levels for rubrics are specifically written to be descriptions of work at various quality levels. Whether the rubric scale is defined as 1, 2, 3, and 4 or as A, B, C, D, and F, or some other scale, assigning a level to a piece of work in this manner is an example of the quality-level method.

Grading a Composite of Several Scores

This section discusses how to combine scores from several components into a single composite mark. The discussion is consistent with the criterion-referenced grading framework. When using a criterion-referenced framework, as with norm-referenced grading, you must be careful when assigning weights to components. If weights are assigned improperly, the composite results will not maintain the importance you seek for each component.

Fixed-Percentage Method If you use a fixed-percentage grading method, you will have a percentage score for each student for each component. Then, you multiply each component percentage by its corresponding weight, add these products together, and divide the sum of products by the sum of the weights. This procedure may be summarized by the following formula:

composite percentage score

$$= \frac{\Sigma(\text{weight} \times \text{percentage score})}{\Sigma(\text{weight})} \qquad \text{[Eq. 14.2]}$$

where

$\sum$ = sum of

weight = weight you give to a component

percentage score = the percentage you gave the student on the component

To illustrate, consider Figure 14.14.

If you did not use the weights, each component would count equally toward the composite. This procedure should not be used with norm-referenced grading because the weights assigned here fail to reflect the standard deviations of the components.

Total Points Method The way we described the total points method in the previous section automatically grades composites. The composite score for a student is the total of the points the student accumulates. However, make sure that the points you assign for each component reflect the weight you want each component to contribute to the total composite. For example, if the weights you want for the components are quizzes 20%, homework 10%, term paper 20%, and exam 50%, then points for each component should reflect these percentages of the total maximum points. Thus, if the maximum total points is 200, then all of the quizzes are worth a maximum of 40 points (= 20% of 200), all of the homework a maximum of 20 points, term paper 40 points, and exam 100 points.

Quality-Level Methods You can derive a grade from a set of rubric scores on various assignments in one of several ways: summing across components, using the median score, or using rules for minimum attainment. These methods may also be used when the components are a mixture of percentage scores on tests and quizzes, and rubrics-based scores. As we pointed out in our discussions of the other methods, be careful to place all of the component marks on comparable scales before combining them into a composite for grade assignment. So for instance, all components marks may be converted into an A, B, C, D, and F quality scale before combining them to arrive at a final grade. These letter grades (as well as rubrics-based marks) represent achievement scales on which students are *partially ordered*.

Using the Median Score This works well for components that include a mixture of rubrics and percent-correct scores. The **median score method** approach treats all component marks as ordinal data (i.e., essentially as ranks) and uses the student's median mark to calculate the grade instead of using the sum of marks or the average mark. Before taking the median, convert all scores (rubrics, percents, and so on) to the same scale (for example, A, B, C, D, F). The median is discussed in Appendix I. See Brookhart (2009) for a more complete explanation of this method.

Using Minimum Attainment The **minimum attainment method** bases the composite grades on whether students meet minimum standards on the most important assessments that comprise the final

FIGURE 14.14 Example of how to calculate the composite score using the fixed-percentage method.

Suppose you had four components (quizzes, homework, term paper, and exam) that you want to combine into a composite score for the end of a marking period. Suppose, further, that each component was originally marked as a percentage correct. Suppose, too, you did not want to weigh each component the same. Finally, suppose that the students' marks and weights for each component were as follows:

Student	*Quizzes (wt. = 20%)*	*Homework (wt. = 10%)*	*Term paper (wt. = 20%)*	*Exam (wt. = 50%)*	*Weighted composite percentage*
Bob	87	85	70	80	80
Chad	85	80	80	70	75
Susan	75	82	85	60	65
Theresa	70	78	75	65	69

You calculate the weighted composite score (last column) and compare that score to the boundaries you set for the letter grades. You use Equation 14.2 to calculate the weighted composite score. The calculations are as follows:

$$\text{weighted composite score for Bob} = [20 \times 87 + 10 \times 85 + 20 \times 70 + 50 \times 80] \div [100] = 80$$
$$\text{weighted composite score for Chad} = [20 \times 85 + 10 \times 80 + 20 \times 80 + 50 \times 70] \div [100] = 75$$
$$\text{weighted composite score for Susan} = [20 \times 75 + 10 \times 82 + 20 \times 85 + 50 \times 60] \div [100] = 65$$
$$\text{weighted composite score for Theresa} = [20 \times 70 + 10 \times 78 + 20 \times 75 + 50 \times 65] \div [100] = 69$$

Suppose your grade boundaries were:

A = 90–100; B = 80–89; C = 70–79; D = 60–69; and F = 0–59

Then using the weighted composite percentages as calculated, the grades for these students are:

Bob = B; Chad = C; Susan = D; and Theresa = D

grade, while at the same time allowing somewhat lower performance on a few of the less important components. Although this method could be used in a variety of circumstances, it is suitable when you have marked the components using quality-level scores such as letter grades (see Figure 14.10), rubric scores (see Figure 12.8), or quality-level labels (e.g., basic, proficient, advanced) but you do not want to convert these quality-level marks to percentages.

The minimum attainment rules method is a *noncompensatory approach to grading*. The methods whereby you add together scores from the components are called *compensatory* methods because a student's low score on one component can be compensated by a high grade on another.

In the minimum attainment method, a teacher sets the minimum marks on some important assessment components that the students must meet in order to receive a particular grade. If students fail to meet the minimum standards on these specified assessments, they cannot receive high grades, no matter how well they did on the other, less important, assessments. Students who *do* meet the minimum standards on the specified assessment also must meet some standards on the other, less important components. The minimum attainment rules method is only one such noncompensatory approach to grading.

To use this method, you first determine what components will be included in students' final grades, and which of those are more important to demonstrating the students' achievement of the learning targets. Second, you must specify, for each of these "more important" components, the minimum level of performance you will accept for each of the final grade levels of A, B, C, D, and so on. Third, you establish rules for what levels of performance you will accept, at each final grade level, on each of the "less important" components. These rules form a set of decision rules for how to assign grades. An example of how to use these rules follows:

Example

Example of the minimum attainment method for grading

Assume an English class with one test (graded in percentages that are then converted to letter grades), four small writing assignments (graded with rubrics as A, B, C, D, F), and one longer paper (also graded with rubrics as A, B, C, D, F). That is, six components go into the final grade. Assume, also, you want the combined test and paper marks to be worth twice as much as the four smaller assignments.

If a student scores	*Then the grade is*
As on at least three of the writing assignments, *and* As on the paper and test, or an A on one and a B on the other	A
As or Bs on at least three of the writing assignments, *and* at least Bs on the paper and test, or an A on one and a C on the other	B
C or better on at least three of the writing assignments, *and* at least Cs on the paper and test, or a B or better on one and a D or better on the other	C
D or better on at least three of the writing assignments, *and* at least Ds on the paper and test, or a C or better on one and an F on the other	D
A combination lower than the above	F

Below an example of how these rules would be applied for eight students. You may notice from the example that the rules are similar to the rules in set theory arithmetic used in elementary schools because they use *if, not, and, or,* and *then*. In the preceding example, for instance, the rule for an A grade states: "IF (writing assignments = 3 As or more) AND [(both paper and test = A) OR (paper and test have A and B)] THEN overall grade = A. Sometimes this method of grading is referred to as the **logic rule method** (Arter & McTighe, 2001).

Of course, you may use other decision rules beside the ones we used in the example. Other decision rules might describe minimum attainment in the manner of an holistic rubric (as in Figure 14.10), for example.

Gradebook Computer Programs

A number of the procedures described for calculating composite grades are somewhat complex and involve some tedious multiplication and addition. All these calculations can be made with the help of a handheld calculator, of course. If you have a personal computer, you may also want to use a simple spreadsheet program to make the calculations. Several **gradebook programs** in the marketplace can also help you. The advantage is that a gradebook program provides you with a spreadsheet already set up for recording and reporting grades. The better programs combine spreadsheets and database functions. These will allow you to choose from a variety of grading frameworks, keep a class roster, keep attendance, record comments about students' assignments, obtain class summaries, and print reports for the total class or for one student to take home.

School districts sometimes provide—and require—teachers to use a particular gradebook program. These programs are sometimes linked to the district's administrative software so that report cards can be printed without the extra step of "turning in grades." Some of these programs are linked to a Website where parents, with password and identification, can log in and check their students' grades at any time, and sometimes even compare their student's grade with the rest of the class. This opens up new opportunities for home-school communication.

Example

Example of applying the minimum attainment method for grading in the preceding example to eight students.

	Writing 1	*Writing 2*	*Writing 3*	*Writing 4*	*Long paper*	*Test*	*Final grade*
Aiden	A	A	C	A	A	A	**A**
Anthony	A	B	A	A	A	B	**A**
Ashley	A	B	B	C	B	B	**B**
Billy	A	B	B	C	B	B	**B**
Blake	C	C	C	A	C	C	**C**
Chad	D	D	D	A	D	D	**D**
Jesse	D	D	D	A	F	C	**D**
Sophia	D	D	F	F	D	D	**F**

It also requires even clearer grading plans and policies, so that students and parents who check incomplete records for a marking period correctly interpret the information in front of them. Smith and Walker (2002) recommend that before implementing a buildingwide electronic gradebook system, principals should consider (a) teachers' technology comfort level, (b) computer availability, (c) network capability, (d) interface with student information management system, (e) staff development, (f) ongoing support, and (g) principal's commitment.

One disadvantage of some gradebook programs is that they limit the type of grading you may employ, or they may not permit you to use your own grading method to override the method(s) built into the program. We have seen a gradebook program advertised that claims to "think like an elementary school teacher" and includes ways to encode "effort" into students' grades! Be a careful and critical consumer of any program you choose. If your district chooses a gradebook program for you, you should still investigate what kind of framework it uses for its calculations and adjust default settings to what you intend for your grades whenever possible.

Software for delivering online courses also includes gradebook capability. If you are teaching online, use the same approach to these gradebooks as you would for a gradebook program you use for a face-to-face class. Find out what its capabilities are, what kinds of data it will accommodate, and how it will display summaries or print reports. Most important, find out what framework it uses for combining individual grades or scores into composite marks and check that the method is what you intend. If not, adjust the program's settings.

CONCLUSION

The main theme of this chapter is that in most situations, it is best to combine measures of classroom achievement to create a grading scale that will communicate achievement information to students and parents. Assess and report citizenship, behavior, and work habits separately. We demonstrated various grading methods; your choices should be based on your teaching philosophy, district grading policy, and classroom context.

This ends our discussion of educational assessment in the classroom. The next section discusses standardized testing.

EXERCISES

1. Prepare a brief paper explaining the grading system you use (or plan to use). In a separate section explain the educational rationale for using this system, including an explanation of how your system has improved (or will improve) your students' educational development. Discuss your grading point of view with others in your class. Prepare at least one paragraph explaining each of the following:
   a. The meaning of your grade symbols.
   b. The meaning of failure in your class.
   c. How you distinguish between "failure" and "failure to try."
   d. How you handle late work or work not handed in.
   e. How you avoid the "deadly zero."
   f. What student performances count toward grades you assign your student.
   g. The number of each letter (or other symbol) grade you typically assign (or will assign) in your class.
   h. What components go (or will go) into the end-of-term grade for your students.
   i. How much weight each component in Item h should receive.
   j. What boundaries you use (or would use) for each grade.
   k. How you handle students who are on the borderline between grades.
   l. Any other factors you take into account.
2. Talk with school administrators and teachers at several grade levels in the school district in which you live or work. Bring Figure 14.5 with you.
   a. What method(s) of student progress reporting is (are) used?
   b. Is the district satisfied with the method(s) it uses?
   c. Which of the advantages and disadvantages listed in Figure 14.5 has the school district experienced? Explain.
   d. Obtain copies of the district's report card(s). Share all your findings with the other members of your class.
   e. Summarize the similarities and differences among the district represented in your class and offer suggestions for improving student progress reporting.

FIGURE 14.15 List of students and the marks they received on each component during one marking period. Use this table for Exercise 4.

Pupil	Last year's grade average	Teacher's judgment of ability	Deportment	Homework				Quizzes		
				1	2	3	Project	1	2	Test score
A	B	Average	Very good	10	3	8	12	8	4	25
B	C	Average	Very good	9	2	7	15	7	4	20
C	A	Very high	Poor	10	0	9	15	10	5	29
D	A	Above average	Excellent	10	4	10	15	6	5	28
E	D	Average	Poor	0	2	5	0	5	3	10
F	B	Average	Good	10	1	2	10	5	3	18
G	C	Below average	Good	10	3	9	8	6	2	15
H	C	Above average	Poor	10	1	4	15	8	4	12
I	C	Above average	Excellent	10	1	3	13	8	3	21
J	C	Above average	Very good	10	1	5	10	8	2	23
Maximum possible score:				10	10	10	15	10	5	30
Mean				8.9	1.8	6.2	11.3	7.1	3.5	20.1
Standard deviation:				3.0	1.2	2.6	4.5	1.5	1.0	6.1
Teacher's weights:				5%	5%	5%	15%	10%	10%	50%

3. Identify a unit that you have taught or will teach.
   a. In the context of your teaching situation and this unit, identify the assessment variables, the reporting variables, and the grading variables.
   b. Prepare a three-column table listing these variables and describing how you have assessed (or will assess) each one.
   c. Share your findings with the others in your class.
4. Figure 14.15 contains information about the performance of a class of 10 students. Use it to complete this exercise.
   a. Determine an overall report card grade for each student using the following methods: (i) self-referencing; (ii) criterion-referencing, fixed-percentage; (iii) criterion-referencing, total points; and (iv) norm-referencing, SS-score (standard deviation) method.
   b. Prepare a table with the students' names as the row headings and the four different methods as the column headings. Enter the students' grades under each method and compare the results.
   c. Share your results with the others in your class. Where do you see the most agreement and most disagreement?
   d. List the reasons for agreements and disagreements for each method.
5. Name several kinds of student performances (homework, class participation, performance tasks, tests, etc.) that you believe should be included in each of the following levels: primary, middle school, or high school.
   a. State what weight should be assigned to each type of performance. Explain the reasons for these weights by discussing each of the six factors stated in the chapter in relation to each kind of performance.
   b. Would the weights vary with different grade levels or with different subjects? Explain.

Standardized Achievement Tests

KEY CONCEPTS

1. Standardized tests are tests for which the procedures, administration, materials, and scoring rules are fixed so that as far as possible the assessment is the same at different times and places.
2. Standardized achievement tests include multilevel survey batteries, multilevel criterion-referenced tests, other multilevel tests for a single curricular area, and single-level tests for one course or subject area.
3. State- or district-mandated tests include state achievement tests customized to state standards, interim or benchmark tests, Response to Intervention (RTI) assessments, early childhood assessments, and English language proficiency tests.
4. Nonstandardized achievement tests do not report empirical evidence of their development, quality, or effectiveness.
5. Standardized test results can be used both within and outside the classroom.
6. Avoid inappropriate use of standardized test results.
7. Follow prescribed administrative procedures when you give standardized tests.
8. Prepare your students for standardized testing in ethical and appropriate ways.

IMPORTANT TERMS

empirically documented tests
generalizability of assessment results
in-level versus out-of-level testing
interim or benchmark assessments
multilevel survey battery versus single-level test
special norms
state-mandated assessments
test levels

From Chapter 15 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

OVERVIEW OF STANDARDIZED TESTS

Standardized tests are tests for which the procedures, administration, materials, and scoring rules are fixed so that as far as possible the assessment is the same at different times and places. Achievement tests vary in their purpose, usefulness, and quality. To appreciate their variety, you may find it helpful to classify them. Here is one classifying scheme:

I. Published achievement tests

   A. *Standardized,* **empirically documented tests** have a high degree of standardization. Standardized tests follow the development procedures outlined in Chapter 17, especially the steps that require using empirical data to document their effectiveness. The following types are in this group:

      1. *Multilevel survey batteries* are the familiar, annually administered tests that survey students' general educational growth or basic skill development in each of several curricular areas. *Multilevel* means that the test content spans several grade levels; *battery* means that several curricular areas are assessed by different subtests.
      2. *Multilevel criterion-referenced tests for a single curricular area* provide detailed information about students' status for a well-defined domain of performance in a single subject area (e.g., mathematics). The test spans several grade levels.
      3. *Other multilevel tests for a single curricular area* are noncriterion-referenced tests that assess students in a broader way than do subtests in a survey battery.
      4. *Single-level standardized tests for one course or subject* are developed for assessing achievement at only one educational level or for one course (e.g., Algebra I). Usually they are stand-alone tests, neither coordinated with tests from other courses nor normed on the same students as other tests.

   B. The *No Child Left Behind Act of 2001* and the *Individuals with Disabilities Education Improvement Act (IDEA) of 2004* have necessitated state compliance with testing and reporting requirements on a much larger scale than ever before. Test publishers have created many assessments to address these requirements. State- or district-mandated tests include state achievement tests customized to state standards, interim or benchmark tests, Response to Intervention (RTI) assessments, early childhood assessments, and English language proficiency tests.

      1. *State-mandated customized tests* are developed by publishers of standardized multilevel survey batteries for use only in a particular state. The tests are said to be *customized* because a publisher contracts with a state to prepare standardized tests that are aligned with the state's standards and are secure so they can be used for accountability purposes. Since the NCLB Act, the grades typically covered are 3 through 12 and the subjects tested are reading, language arts, mathematics, and science.
      2. ***Interim*** *or* ***benchmark assessments*** provide information on groups of students and identify, usually in 6- to 9-week cycles, where students are with respect to achievement of state standards to date. Interim assessments can be customized, but many districts also use off-the-shelf interim tests.
      3. *Response to Intervention (RTI) assessments* focus on early identification of struggling learners and the delivery of targeted interventions. Most RTI solutions from test publishers have a variety of assessment and intervention tools and employ a tiered approach which progresses from universal screening to progress monitoring and interventions.
      4. *Early childhood* assessments are of several types. Some are designed to assess academic readiness of concepts (e.g., colors, letters, numbers, sizes) that are seen to be directly related to early childhood education or predict readiness for more formal education (Bracken, 2002). Others are designed to assess student performance like the interim or benchmark assessment described above, but with the focus on Pre-K to grade 2 students.
      5. The purpose of *English language proficiency* assessments is to place English language learners (ELLs) at the appropriate proficiency level for bilingual or English as a second language programs, or determine

if ELLs are ready to exit these programs. ELL testing is required under Title III: Language Instruction for Limited English Proficient and Immigrant Students of NCLB.

C. *Nonstandardized tests, without adequate empirical data to document* their effectiveness, make little or no attempt to standardize and do not follow all the development procedures outlined in Chapter 17. Publishers do not spend the time and money to document their effectiveness or their quality empirically.
   1. *Some criterion-referenced tests* estimate students' status with respect to a well-defined domain of performance (usually specified by specific behavioral objectives), but they lack standardization and empirical documentation of worth.
   2. *Textbook or curricular accompaniments* are tests or test items found in teacher's editions, at the end of textbook chapters, at the back of the book, in supplements that come with textbook series, or built into instructional materials. They are called different names, such as pretests, posttests, placement tests, progress checks, unit tests, review tests, or curriculum-embedded tests. They usually lack standardization and empirical data to document their quality. These tests often measure low-level cognitive skills, and sometimes have several incorrectly keyed answers.

II. *Teacher-made tests* are tests you create to measure the specific learning targets your curriculum framework emphasizes. These tests help you in making day-to-day instructional decisions.

The focus of this chapter is limited to standardized tests having empirical data to document their effectiveness. Standardizing is necessary if you want the results to be comparable from time to time, place to place, and person to person. If an assessment procedure is standardized, you are better able to properly interpret students' scores on it. The quality of any assessment procedure is demonstrated by using empirical data to document its validity and effectiveness. These data provide the test developers with a basis for (a) improving and selecting tasks, (b) establishing reliability and validity, (c) describing how well the assessment works in the target population of students, (d) creating scales to measure growth, (e) equating scores (making scores comparable from grade to grade and from one form of the assessment to another), and (f) developing a variety of norm-referenced scores.

TRADITIONAL STANDARDIZED ACHIEVEMENT TESTS

Multilevel Survey Batteries

The workhorse of standardized achievement testing is the **multilevel survey battery**. Although each publisher's test battery emphasizes different details of content and skill, the batteries are organized similarly.

Common Features Most group-administered survey batteries have the following features in common (Iwanicki, 1980).

1. *Test development features.* Manuals and other materials describe for each subtest the (a) content and learning targets covered, (b) types of norms and how they were developed, (c) type of criterion-referencing provided, (d) reliability data, and (e) techniques used to screen items for offensiveness and possible gender, ethnic, and racial bias.
2. *Test administration features.* Tests generally (a) have two equivalent forms; (b) require a total administration time of 2 to 3 hours, spread among several testing sessions over several school days (although tests vary widely in length and administration time); (c) provide practice booklets for students to use before being tested; (d) have separate, machine-scorable answer sheets for upper grades (students in lower grades mark answers directly on the machine-scorable test booklets); and (e) permit both **in-level** and **out-of-level testing.**[1]
3. *Test norming features.* Tests generally use broadly representative national sampling for norms development and provide both fall and spring individual student norms. Sometimes **special norms** such as the following are provided: (a) large-city norms, (b) norms for students in special government entitlement programs, (c) norms

[1] Test booklets are organized by level; each level is designed for use with a few grades. A student is said to be tested *in-level* if the test booklet level corresponds to the student's actual grade placement. If a student's level of academic functioning is either above or below the actual grade placement, the school may administer the test level that more nearly corresponds to the student's functioning level. This is called *out-of-level testing*. A student is measured best when a test is tailored to the student's functioning level.

for high-income communities, (d) norms for nonpublic schools, (e) regional norms, and (f) norms for school-building averages.

4. *Test score features.* Tests provide raw scores for each subtest and the following norm-referenced scores: percentile ranks, normal curve equivalents, stanines, extended normalized standard scores, and grade equivalents (or some similar grade-level indicator score). Attitudes toward using grade equivalents vary. Some tests provide instructional reading-level scores that are keyed to commonly used basal readers. (We discuss these scores in the next chapter.)
5. *Test score reporting and interpretation features.* Tests generally have interpretive manuals for teachers, school administrators, and/or counselors. Most group tests provide computer-prepared narrative reports that contain summaries of district, school-building, and classroom test results. Some types of reports may be provided with the purchase of the test, and others may be sold separately. Figures 15.1 to 15.3 show reports that you would expect to be able to read and interpret for your students.

Differences Although survey batteries share common features, they are definitely not interchangeable. Scores obtained from different publishers' batteries, even on subtests with similar-sounding titles, will be different and cannot be compared directly. Among the features that are different and that seriously affect comparability of scores are the following:

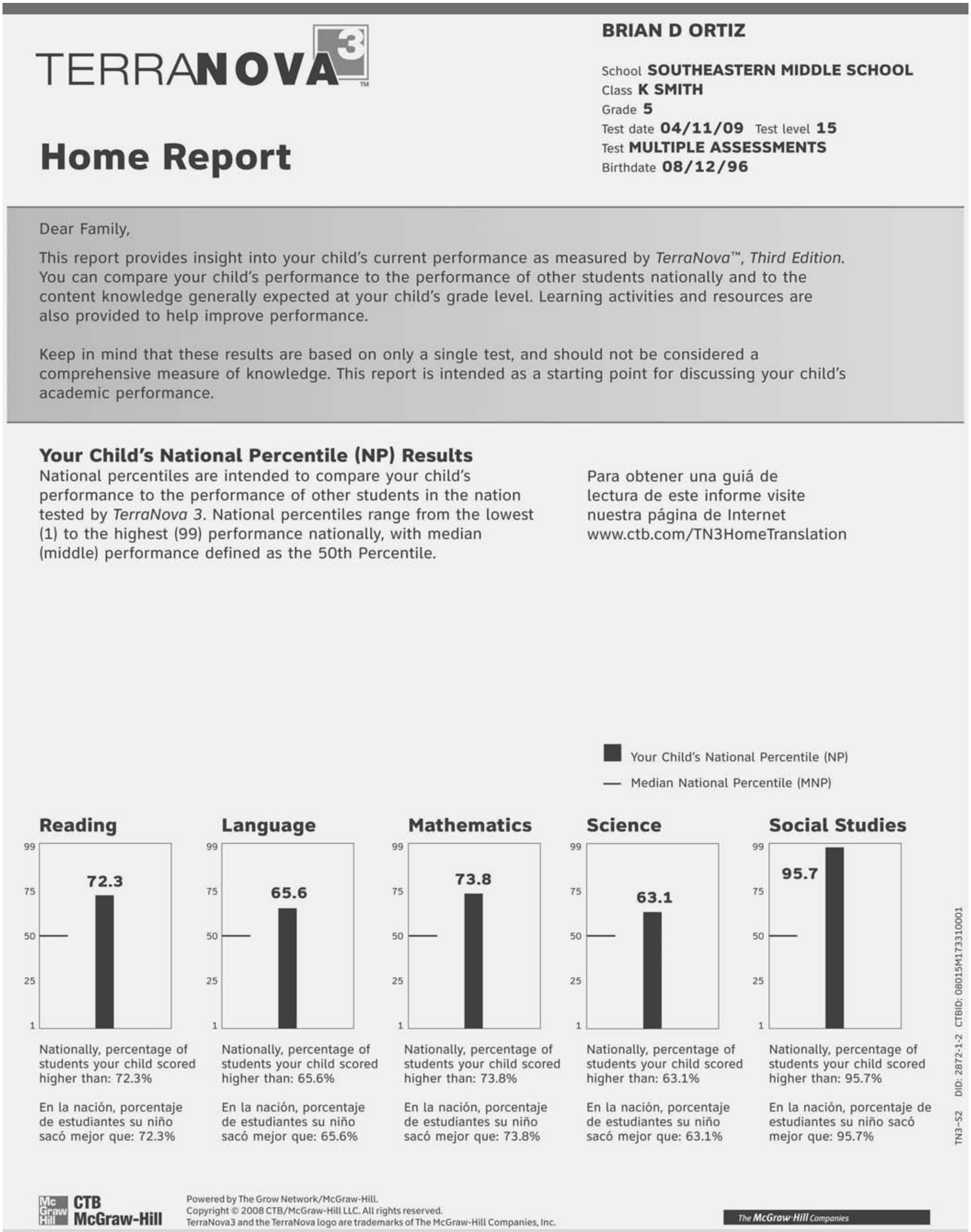
1. *Emphasis within content areas.* Subtest scores on batteries from different test publishers have different meanings. For example, a study of the mathematics subtests of four standardized survey batteries for the fourth-grade level indicated that the percentage of items covering a topic such as fractions varied widely among tests—from 5.4% to 14.4% (Freeman, Kuhs, Knappen, & Porter, 1982). This difference in coverage affects pupils' scores significantly. Because each test publisher chooses to emphasize each subtopic somewhat differently, there may be a serious mismatch between what a given battery calls "reading comprehension" or "social studies" and what a given school and/or teacher emphasizes in class. These mismatches result in subtest scores (e.g., grade equivalents in a subject area) that may not represent a student's current level of functioning. The overlap of a test and a school's instructional program is an extremely important consideration in choosing among test batteries. The importance of aligning accountability tests with state standards was discussed in Chapter 3.

2. *Quality of developmental scales' articulation between grade levels.* One use of a standardized test is to measure students' growth on a continuous scale. If the scale is constructed properly, it is possible to track students' educational growth over the various grade levels. But different test developers use different technical methods for creating developmental scales, even when the scales have the same name. A chief consideration in the practical use of test results concerns the amount of grade-to-grade overlap in the development scores (Peterson, Kolen, & Hoover, 1989). For example, a fourth grader may have a grade-equivalent of 6.0 on the fourth-grade mathematics subtest and a grade-equivalent of 4.9 on the sixth-grade counterpart of the same subtest. Different techniques for constructing grade equivalents will create differing amounts of overlap. If this happens, the result will be scores that show a spuriously erratic pattern of growth for youngsters as they progress through the grades.

3. *Quality of services offered to schools.* Test publishers differ in the extent of their technical support and interpretative services for schools using their products. Some publishers sell the product and certain standard services (such as computer printouts summarizing test results for a school district) but do not provide knowledgeable consultants who can advise a school on particular problems or even on how to interpret their results in general. A school official who is planning to purchase a survey battery should explore fully with the publisher's sales representative the nature and cost of technical support services that will come with the test battery.

Organization of Batteries Each battery is group administered and each contains several subtests. A subtest assesses one area, such as reading, mathematics, listening skills, English usage (mechanics), writing, spelling (recognition), vocabulary (word meaning), or skills in using library and reference materials. Not all questions on these subtests are multiple-choice: In recent years, publishers have added constructed-response items or performance tasks to several subtests or have offered

FIGURE 15.1 **Example of the first page of a computer-prepared narrative report on an individual student's standardized test performance. The report is meant to be sent home to parents.**



TERRANOVA 3™

Home Report

BRIAN D ORTIZ

School **SOUTHEASTERN MIDDLE SCHOOL**
Class **K SMITH**
Grade **5**
Test date **04/11/09** Test level **15**
Test **MULTIPLE ASSESSMENTS**
Birthdate **08/12/96**

Dear Family,

This report provides insight into your child's current performance as measured by *TerraNova™, Third Edition*. You can compare your child's performance to the performance of other students nationally and to the content knowledge generally expected at your child's grade level. Learning activities and resources are also provided to help improve performance.

Keep in mind that these results are based on only a single test, and should not be considered a comprehensive measure of knowledge. This report is intended as a starting point for discussing your child's academic performance.

Your Child's National Percentile (NP) Results

National percentiles are intended to compare your child's performance to the performance of other students in the nation tested by *TerraNova 3*. National percentiles range from the lowest (1) to the highest (99) performance nationally, with median (middle) performance defined as the 50th Percentile.

Para obtener una guiá de lectura de este informe visite nuestra página de Internet www.ctb.com/TN3HomeTranslation

■ Your Child's National Percentile (NP)

— Median National Percentile (MNP)

Reading

Nationally, percentage of students your child scored higher than: 72.3%

En la nación, porcentaje de estudiantes su niño sacó mejor que: 72.3%

Language

Nationally, percentage of students your child scored higher than: 65.6%

En la nación, porcentaje de estudiantes su niño sacó mejor que: 65.6%

Mathematics

Nationally, percentage of students your child scored higher than: 73.8%

En la nación, porcentaje de estudiantes su niño sacó mejor que: 73.8%

Science

Nationally, percentage of students your child scored higher than: 63.1%

En la nación, porcentaje de estudiantes su niño sacó mejor que: 63.1%

Social Studies

Nationally, percentage of students your child scored higher than: 95.7%

En la nación, porcentaje de estudiantes su niño sacó mejor que: 95.7%

TN3-S2 DID: 2872-1-2 CTBID: 08015M173310001

CTB McGraw-Hill

Powered by The Grow Network/McGraw-Hill.
Copyright © 2008 CTB/McGraw-Hill LLC. All rights reserved.
TerraNova3 and the TerraNova logo are trademarks of The McGraw-Hill Companies, Inc.

The McGraw-Hill Companies

Source: From *Sample Home Report, TerraNova,* 3rd Edition. Copyright © 2008 by CTB/McGraw-Hill. Reproduced with permission of The McGraw-Hill Companies, Inc.

FIGURE 15.2 Example of a report that analyzes your class's performance on clusters of items. This report is for mathematics subtests. It reports the percent correct (%C) for each item and for clusters of items. Comparisons are made for your class, the district at your grade level, and the nation at your grade level. This type of report helps you see where in the mathematics area your class is strong and weak.

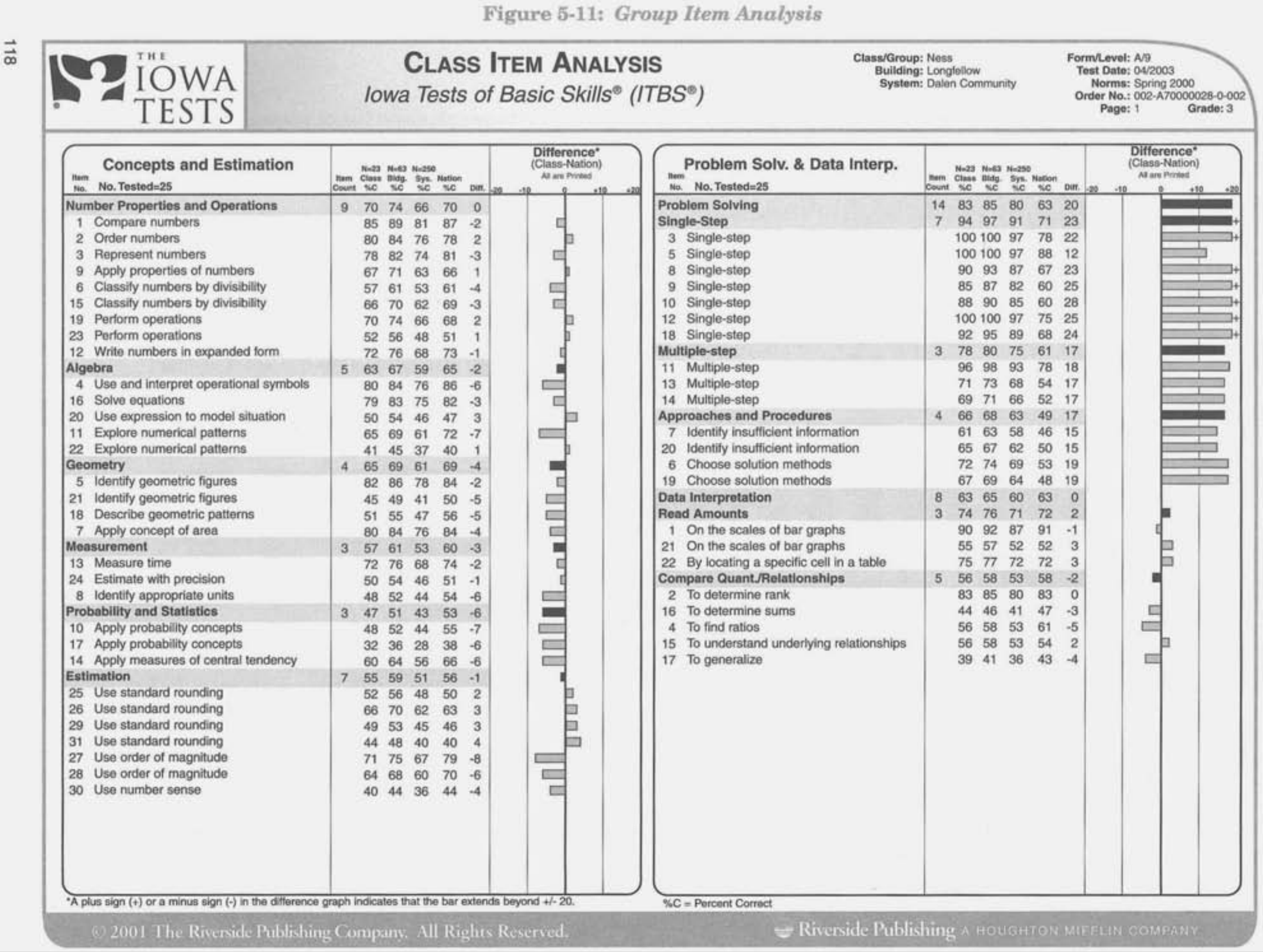


Figure 5-11: *Group Item Analysis*

118

THE IOWA TESTS

Class Item Analysis

Iowa Tests of Basic Skills® (ITBS®)

Class/Group: Ness
Building: Longfellow
System: Dalen Community

Form/Level: A/9
Test Date: 04/2003
Norms: Spring 2000
Order No.: 002-A70000028-0-002
Page: 1 **Grade:** 3

Concepts and Estimation

Item No.	No. Tested=25	Item Count	N=23 Class %C	N=63 Bldg. %C	N=250 Sys. %C	Nation %C	Diff.
	Number Properties and Operations	9	70	74	66	70	0
1	Compare numbers		85	89	81	87	-2
2	Order numbers		80	84	76	78	2
3	Represent numbers		78	82	74	81	-3
9	Apply properties of numbers		67	71	63	66	1
6	Classify numbers by divisibility		57	61	53	61	-4
15	Classify numbers by divisibility		66	70	62	69	-3
19	Perform operations		70	74	66	68	2
23	Perform operations		52	56	48	51	1
12	Write numbers in expanded form		72	76	68	73	-1
	Algebra	5	63	67	59	65	-2
4	Use and interpret operational symbols		80	84	76	86	-6
16	Solve equations		79	83	75	82	-3
20	Use expression to model situation		50	54	46	47	3
11	Explore numerical patterns		65	69	61	72	-7
22	Explore numerical patterns		41	45	37	40	1
	Geometry	4	65	69	61	69	-4
5	Identify geometric figures		82	86	78	84	-2
21	Identify geometric figures		45	49	41	50	-5
18	Describe geometric patterns		51	55	47	56	-5
7	Apply concept of area		80	84	76	84	-4
	Measurement	3	57	61	53	60	-3
13	Measure time		72	76	68	74	-2
24	Estimate with precision		50	54	46	51	-1
8	Identify appropriate units		48	52	44	54	-6
	Probability and Statistics	3	47	51	43	53	-6
10	Apply probability concepts		48	52	44	55	-7
17	Apply probability concepts		32	36	28	38	-6
14	Apply measures of central tendency		60	64	56	66	-6
	Estimation	7	55	59	51	56	-1
25	Use standard rounding		52	56	48	50	2
26	Use standard rounding		66	70	62	63	3
29	Use standard rounding		49	53	45	46	3
31	Use standard rounding		44	48	40	40	4
27	Use order of magnitude		71	75	67	79	-8
28	Use order of magnitude		64	68	60	70	-6
30	Use number sense		40	44	36	44	-4

Problem Solv. & Data Interp.

Item No.	No. Tested=25	Item Count	N=23 Class %C	N=63 Bldg. %C	N=250 Sys. %C	Nation %C	Diff.
	Problem Solving	14	83	85	80	63	20
	Single-Step	7	94	97	91	71	23
3	Single-step		100	100	97	78	22
5	Single-step		100	100	97	88	12
8	Single-step		90	93	87	67	23
9	Single-step		85	87	82	60	25
10	Single-step		88	90	85	60	28
12	Single-step		100	100	97	75	25
18	Single-step		92	95	89	68	24
	Multiple-step	3	78	80	75	61	17
11	Multiple-step		96	98	93	78	18
13	Multiple-step		71	73	68	54	17
14	Multiple-step		69	71	66	52	17
	Approaches and Procedures	4	66	68	63	49	17
7	Identify insufficient information		61	63	58	46	15
20	Identify insufficient information		65	67	62	50	15
6	Choose solution methods		72	74	69	53	19
19	Choose solution methods		67	69	64	48	19
	Data Interpretation	8	63	65	60	63	0
	Read Amounts	3	74	76	71	72	2
1	On the scales of bar graphs		90	92	87	91	-1
21	On the scales of bar graphs		55	57	52	52	3
22	By locating a specific cell in a table		75	77	72	72	3
	Compare Quant./Relationships	5	56	58	53	58	-2
2	To determine rank		83	85	80	83	0
16	To determine sums		44	46	41	47	-3
4	To find ratios		56	58	53	61	-5
15	To understand underlying relationships		56	58	53	54	2
17	To generalize		39	41	36	43	-4

*A plus sign (+) or a minus sign (-) in the difference graph indicates that the bar extends beyond +/- 20.

%C = Percent Correct

Riverside Publishing A HOUGHTON MIFFLIN COMPANY

Source: S. Dunbar, H. D. Hoover, D. A. Frisbie, and K. R. Oberley, 2008. Copyright © 2008 by The University of Iowa. All rights reserved. Reproduced from the *Iowa Tests®, Interpretive Guide for School Administrators*, p. 118, with permission of the Riverside Publishing Company.

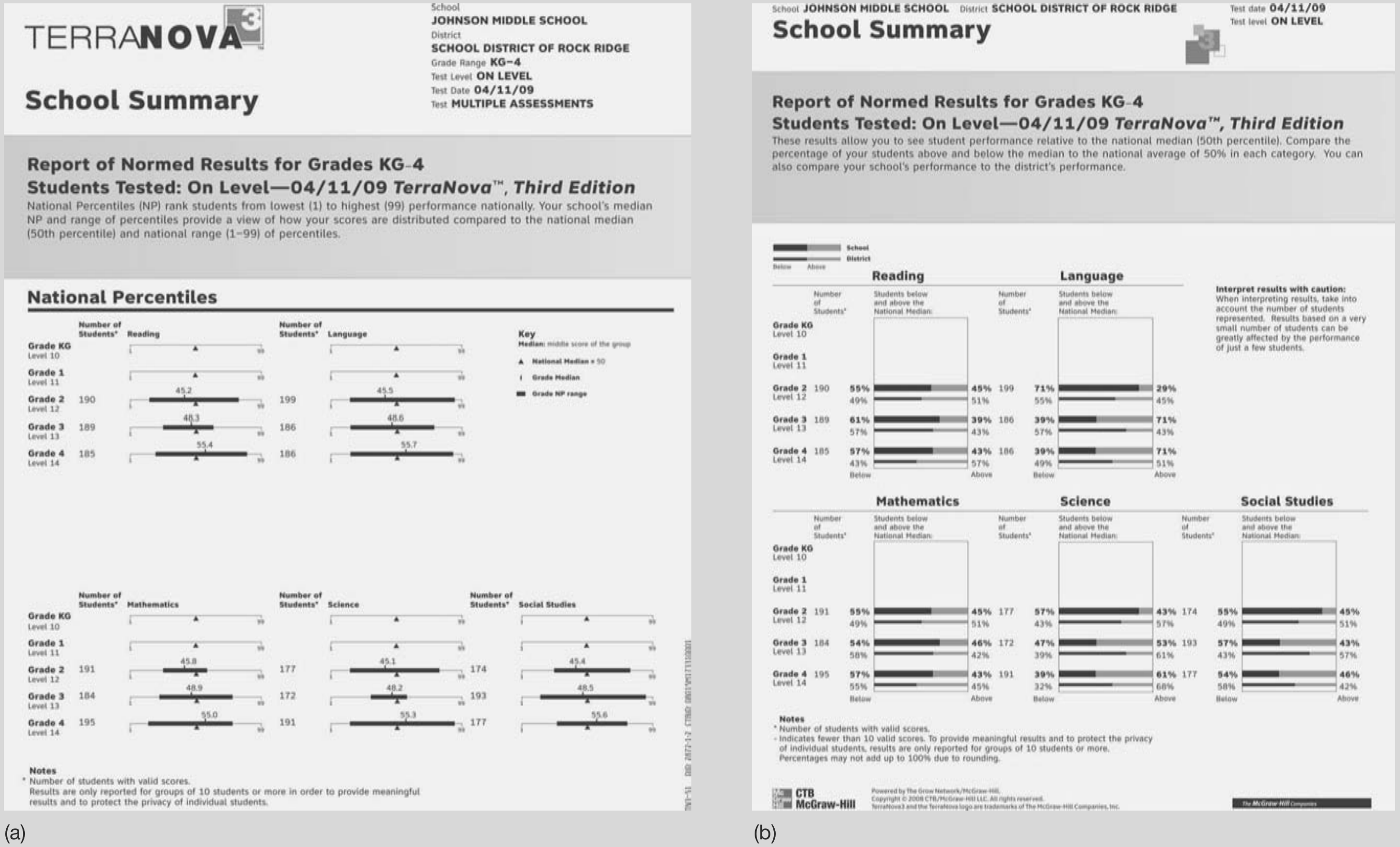
them as separate subtests. Separate scores are given for each subtest. Usually, a battery has subtests for six to eight curriculum areas. Different publishers may have different subtest names for the same curriculum area.

Each subtest is made up of a coordinated series of **test levels** that spans the grades. For example, a reading subtest may be organized into four levels: one level for Grades 1 and 2, another for Grades 3 and 4, another for 5 and 6, and a fourth for 7 and 8. It is not unusual for a publisher to have adjacent levels with overlapping grades (e.g., one level covering Grades 3–4–5 and the next level covering Grades 5–6–7).

Many achievement survey batteries are available in the marketplace. (Each edition of the *Mental Measurements Yearbook* lists dozens of different tests.) Here, in alphabetical order, are the most widely used batteries. The grade ranges they cover are noted in parentheses.

Iowa Tests of Basic Skills (K–8)
Iowa Tests of Educational Development (9–12)
Peabody Individual Achievement Test (K–adult)
Stanford Achievement Test Series (K–12)
TerraNova, Third Edition, (K–12)
Wide Range Achievement Test (K–adult)

FIGURE 15.3 Example of a building report showing the performance of a school's Grades 2 through 4 students on each subtest. This report shows (a) how the students in each grade performed relative to the national norm group; and (b) the percent of students in each grade in the school and in the district who scored above and below the national median (the 50th percentile).



TERRANOVA 3

School Summary

School
JOHNSON MIDDLE SCHOOL
District
SCHOOL DISTRICT OF ROCK RIDGE
Grade Range KG–4
Test Level ON LEVEL
Test Date 04/11/09
Test MULTIPLE ASSESSMENTS

Report of Normed Results for Grades KG–4
Students Tested: On Level—04/11/09 *TerraNova™*, *Third Edition*

National Percentiles (NP) rank students from lowest (1) to highest (99) performance nationally. Your school's median NP and range of percentiles provide a view of how your scores are distributed compared to the national median (50th percentile) and national range (1–99) of percentiles.

National Percentiles

	Number of Students*	Reading	Number of Students*	Language
Grade KG Level 10				
Grade 1 Level 11				
Grade 2 Level 12	190	45.2	199	45.5
Grade 3 Level 13	189	48.3	186	48.6
Grade 4 Level 14	185	55.4	186	55.7

Key
Median: middle score of the group
▲ National Median = 50
I Grade Median
▬ Grade NP range

	Number of Students*	Mathematics	Number of Students*	Science	Number of Students*	Social Studies
Grade KG Level 10						
Grade 1 Level 11						
Grade 2 Level 12	191	45.8	177	45.1	174	45.4
Grade 3 Level 13	184	48.9	172	48.2	193	48.5
Grade 4 Level 14	195	55.0	191	55.3	177	55.6

Notes
* Number of students with valid scores.
Results are only reported for groups of 10 students or more in order to provide meaningful results and to protect the privacy of individual students.

(a)

School JOHNSON MIDDLE SCHOOL District SCHOOL DISTRICT OF ROCK RIDGE
Test date 04/11/09
Test level ON LEVEL

School Summary

Report of Normed Results for Grades KG–4
Students Tested: On Level—04/11/09 *TerraNova™*, *Third Edition*

These results allow you to see student performance relative to the national median (50th percentile). Compare the percentage of your students above and below the median to the national average of 50% in each category. You can also compare your school's performance to the district's performance.

▬ School
▬ District
Below Above

	Reading: Number of Students*	Reading: Below	Reading: Above	Language: Number of Students*	Language: Below	Language: Above
Grade KG Level 10						
Grade 1 Level 11						
Grade 2 Level 12	190	55% 49%	45% 51%	199	71% 55%	29% 45%
Grade 3 Level 13	189	61% 57%	39% 43%	186	39% 57%	71% 43%
Grade 4 Level 14	185	57% 43%	43% 57%	186	39% 49%	71% 51%

Students below and above the National Median:

Interpret results with caution: When interpreting results, take into account the number of students represented. Results based on a very small number of students can be greatly affected by the performance of just a few students.

	Mathematics: Number of Students*	Mathematics: Below	Mathematics: Above	Science: Number of Students*	Science: Below	Science: Above	Social Studies: Number of Students*	Social Studies: Below	Social Studies: Above
Grade KG Level 10									
Grade 1 Level 11									
Grade 2 Level 12	191	55% 49%	45% 51%	177	57% 43%	43% 57%	174	55% 49%	45% 51%
Grade 3 Level 13	184	54% 58%	46% 42%	172	47% 39%	53% 61%	193	57% 43%	43% 57%
Grade 4 Level 14	195	57% 55%	43% 45%	191	39% 32%	61% 68%	177	54% 58%	46% 42%

Notes
* Number of students with valid scores.
- Indicates fewer than 10 valid scores. To provide meaningful results and to protect the privacy of individual students, results are only reported for groups of 10 students or more.
Percentages may not add up to 100% due to rounding.

CTB McGraw-Hill
Powered by The Grow Network/McGraw-Hill.
Copyright © 2008 CTB/McGraw-Hill LLC. All rights reserved.
TerraNova3 and the TerraNova logo are trademarks of The McGraw-Hill Companies, Inc.
The McGraw-Hill Companies

(b)

Source: From *Score Reports for TerraNova,* 3rd ed. Copyright © 2008 by CTB/McGraw-Hill. Reproduced with permission of The McGraw-Hill Companies, Inc.

These are all group-administered tests, except for the *Peabody* and the *Wide Range*, which are individually administered. Publishers are listed in Appendixes K and L. Details about each battery can be obtained from the publishers' catalogs and Websites. Critical reviews are found in the *Mental Measurements Yearbook*s, *Test Critiques*, and other sources identified in Chapter 17. Figure 15.4 shows the curriculum areas, subtests, and grade levels covered by some of the more popular standardized achievement tests.

Although different publishers' survey batteries are similar in their surface features, they are not interchangeable, even though subtest names may

FIGURE 15.4 **Examples of curriculum areas and grade levels assessed by survey batteries.**

Curriculum area/ subarea[a]	Stanford Achievement (10th Test ed.)	Metropolitan Achievement Tests (8th ed.)	Iowa Tests of Basic Skills (Form C)	Iowa Tests of Educational Development (Form C)	TerraNova3 Complete Battery Plus
Reading multiple-choice					
Alphabet knowledge	K.0–1.5	K.0–K.5	K.1–1.9		
Word/sentence reading	K.0–2.5	1.5–4.5	K.8–3.5		1.0–4.2
Phonetic/structural analysis	1.5–3.5	K.0–4.5	K.1–3.9		1.0–4.2
Decoding skills	K.0–1.5	K.0–1.5	K.1-3.9		1.0–4.2
Vocabulary	2.5–12.9	1.5–12.9	K.1–9.9	9.0–12.9	K–12.9
Comprehension	1.5–12.9	1.5–12.9	K.8–9.9	9.0–12.9	K–12.9
Reading performance assessment	1.5–12.9[b]	1.5–12.9[b]			3.0–12.9[c]
Language multiple-choice					
Punctuation	1.5–12.9	1.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Capitalization	1.5–12.9	1.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Usage	1.5–12.9	1.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Listening	K.0–9.9	K.0–3.5	K.1–9.9		K.6–2.6
Sentence/paragraph organization	1.5–12.9	3.0–8.9	3.0–9.9	9.0–12.9	1.6–12.9
Language/writing performance assessment	3.5–12.9[b]	1.5–12.9[b]	3.0–12.9[b]	3.0–12.9[b]	3.0–12.9c
Spelling multiple-choice	1.5–12.9	1.5–12.9	1.7–9.9	9.0–12.9	2.0–12.9
Mathematics multiple-choice					
Computation	K.0–12.9	K.5–9.5	1.7–9.9	9.0–12.9	K.6–12.9
Concepts	K.0–12.9	K.5–12.9	K.1–9.9	9.0–12.9	K.0–12.9
Problem solving	K.5–12.9	1.5–12.9	K.1–9.9	9.0–12.9	K.6–12.9
Mathematics performance assessment	1.5–12.9[b]	1.5–12.9[b]			3.0–12.9[c]
Study skills multiple-choice					
Maps, graphs, tables	4.5–12.9	3.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Library/reference materials	4.5–12.9	3.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Study skills performance assessment		K.0–8.9[b]			
Science multiple-choice	K.0–12.9	1.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Science performance assessment	1.5–12.9[b]	1.5–12.9[b]			
Social studies multiple-choice	3.5–12.9	1.5–12.9	1.7–9.9	9.0–12.9	1.6–12.9
Social studies performance assessment	1.5–12.9[b]	1.5–12.9[b]			

Notes: [a] Publishers may have somewhat different names for these areas than those used here. Separate scores are not provided for every area.

[b] Assessments in these areas are available as supplements or additional purchase components that are not part of the battery itself.

[c] Part of the Multiple Assessments Edition.

sound similar. The specific content emphasized, the cognitive skills students are required to use to complete the tasks, and the way the norms and scales are developed will be very different from publisher to publisher.

Tests vary in how well they match any school district's curriculum or state's standards. In some curricula, such as reading and perhaps mathematics, the curricula differ very little from one school district to another within a state. The tests and these curricula may match closely. In other curricula such as science and social studies, especially among elementary schools, there are much larger variations between school districts. For a teacher this means that the different subtests in the battery have less value in assessing the specifics of what the teacher taught during the year. However, such subtests can assess general information and general ability to apply knowledge and skill.

These differences make it necessary for school officials to actually inspect the test items before they adopt a battery, matching their local curriculum to the battery's content and skills emphasis. If there is a wide gap between your local curriculum's learning targets and the battery's tasks, do not adopt the survey battery.

Publishers think of each subtest (e.g., reading comprehension) as assessing a continuous dimension that grows or develops over a range of grades. Because each subtest is a graded series of assessments, the publisher can use empirical data to link the levels together and to place the scores of students from every grade on one numerical scale that spans all the grades. This allows you to use a multilevel subtest to measure a student's year-to-year educational development and growth in a curricular area. Different types of educational development scales are explained in Chapter 16.

Each publisher norms and standardizes its tests on different samples of students, so the samples and the resulting norm-referenced scores are not comparable. However, all the subtests in one publisher's survey battery are administered to the same national sample of students. The major advantage of administering all subtests to the same students is that the different subtest results can be referenced to the same norm group, allowing you to compare a student's relative strengths and weaknesses across the different curricular areas. You can assess these strengths and weaknesses, however, only by comparing a student's percentile rank in one curricular area to that student's percentile rank in another. An example of the kind of comparison you make follows:

Example

Shanna is better in mathematics than she is in social studies because her score in mathematics is higher than 98% of the students at her grade level, whereas her score in social studies is higher than only 60% of students at her grade level.

Survey batteries report grade-equivalent scores and standard scores, too, but you should not use them to compare a student's achievement in two curricular areas. Percentile ranks, standard scores, and grade-equivalent scores are explained in Chapter 16.

Common Learning Targets Virtually all published standardized tests cover content and learning targets judged to be common to many schools rather than one specific school district. Therefore, standardized achievement tests are not focused on the teaching emphasis of one teacher, one school, one textbook, or one set of curricular materials. This is an advantage because it gives you an "external" or "objective" view of what your students have learned. It is also a disadvantage because the cognitive skills and knowledge assessed by the test may not have been taught to the students before they were tested. Therefore, it is imperative that a school district carefully compares a test's content and *when* that content is taught in their schools, item by item, to the state's standards and the school district's curriculum framework before deciding to adopt it. Sometimes as few as three or four misaligned items can have a serious impact on the results. Also, a teacher must develop and use his or her own assessment procedures for day-to-day instructional decisions (e.g., whether a student has mastered a specific concept).

Auxiliary Materials Most publishers of standardized, empirically documented tests provide auxiliary materials to help you interpret and use the assessment results. Teacher's manuals describe in considerable detail the intended purpose and uses of the results, often suggesting ways to improve students' skills by using assessment results for instructional planning. Some publishers provide separate manuals for curriculum coordinators and school administrators to help them use assessment results in curriculum evaluation and reports to the school board. Most publishers provide nicely printed score reports that the school district may use both within the school and with students and parents.

Survey Achievement Battery Selection Examine and review each test individually to judge its appropriateness for your purposes. Before selecting an elementary school survey battery, consider these four points:

1. Survey batteries measure only part of the outcomes desired for elementary schools. Use additional assessment procedures to evaluate the other outcomes.
2. Specific content in subjects such as social studies and science may quickly become dated. Tests designed to measure broad cognitive skills or levels of educational development become dated less quickly.
3. Tests measuring broad cognitive skills or levels of educational development need to be supplemented by teacher-made or standardized tests of specific content.
4. Each battery has a different mix and emphasis of content and skills; each is accompanied by various kinds of interpretive aids. Examine a test battery carefully before deciding to purchase it.

Because high school curricula vary so much, choosing a survey battery for this educational level is difficult. School officials should keep the following six points in mind before selecting a high school test battery:

1. Survey batteries that emphasize basic skills (reading, mathematics, language) may be more useful as measures of high school readiness than as measures of high school outcomes (unless a high school program is especially directed toward basic skills development).
2. Some tests are more oriented toward testing specific content than educational development broadly defined. If you want a content-oriented test, review each item on the test carefully to see if the test measures what the school intends.
3. Tests stressing the measurement of levels of educational development that cut across several subject areas rather than knowledge of specific content tend to measure more complex skills and global processes.
4. The variety of course offerings at the high school level makes it more necessary than at the elementary level to examine the content of each survey battery carefully.
5. You may find it necessary to supplement a high school survey battery with assessments measuring content knowledge of specific subjects.
6. A practical consideration is the continuity of measurement from elementary to secondary levels. This often means purchasing a high school battery from the same company that published the elementary school battery.

Complementing Your State Assessment If your state mandates its own assessment, you will need to take its coverage into account before choosing a published standardized test. Most state assessments have accountability as their main purpose. This is not the case for a published standardized test, which is used primarily to measure individual students' educational growth. Keep the following four points in mind if you are trying to select a standardized multilevel achievement test when you are also faced with a state-mandated assessment:

1. All things being equal, choose a standardized test that requires students to demonstrate learning that is very consistent with your state's standards or curriculum framework.
2. If your community does not like the focus of your state-mandated assessment, choose a multilevel achievement test that reflects the community's concerns. For example, your community may not wish to limit assessment to the higher-order thinking and complex problem solving on which the state assessment focuses. The community may wish to know whether basic skills such as computation, reading comprehension, English writing mechanics, and spelling are being learned.
3. Plan to use the chosen test over a period of at least 5 years, so that you can track changes in your school district.
4. Test at grade levels not tested by the state-mandated assessment to avoid overburdening students and teachers.

Individually Administered Surveys Individually administered achievement batteries are commonly used for students with special needs, such as students with disabilities who otherwise would have difficulty taking assessments in group settings. Students who cannot be assessed in groups often can be validly assessed in individual sessions where the assessment administrator can provide the special accommodations they need and can establish greater rapport than is possible in a group. (See Chapter 5, Figure 5.2, for examples of ways to modify tests.)

Sometimes individual achievement batteries are used as "screening" tests to identify students with learning difficulties, or as part of a broader series of individual assessments when a school psychologist conducts a general psychological evaluation. A school district may use individual achievement survey batteries to assess the general educational development of a newly transferred student, or as a double-check on a previously administered group survey test when the results are being questioned for a particular student. Because both the content and norms of an individual assessment are different from the group test, you should proceed very cautiously when double-checking. You can expect a student's results from the two types of tests to correspond only very roughly.

Two commonly administered individual survey achievement tests are the *Wide Range Achievement Test, Third Edition (WRAT-3)* and the *Peabody Individual Achievement Test—Revised-Normative Update (PIAT-R/NU)*. These single instruments contain items that span many ages or grades (essentially ages 5 to adult). Thus, by their very nature they contain few items specifically associated with a given age or grade level. Such tests do not have as much in-depth coverage as group survey tests that have separate levels for each age or grade level. This comment is not necessarily a criticism of these tests. These wide-range tests make a quick assessment of a student's strengths in several basic curricular areas. This quickly obtained assessment helps the teacher determine relatively weak areas needing more in-depth diagnostic follow-up.

The *PIAT-R/NU*'s items are printed on a small easel. Students do not write responses to the multiple-choice items; they must only say or point to the option. Within each subtest the items are arranged in order of difficulty. A student does not take each item; a starting point (called a basal level) and an ending point (called a ceiling level) are established, based on the student's pattern of correct answers and errors.

Multilevel Criterion-Referenced Tests

Multilevel criterion-referenced tests provide information about students' status with respect to the specific learning targets in a domain. Although some survey batteries also provide this information, most surveys assess very broadly or globally defined educational development. Multilevel criterion-referenced tests tend to focus on a more narrowly defined set of learning targets. Some publishers make efforts to align their tests with states' standards.

Other Multilevel Tests

Other types of multilevel tests are stand-alone products that cover one curricular area, such as reading or mathematics, across several grades. These assessments provide a deeper and broader sampling of content than a corresponding subtest of a survey battery. Thus, more time is devoted to assessing students in a single curricular area than when you use a survey battery subtest. However, if the same sample of students was not used to norm a stand-alone multilevel test concurrently with tests from other curricular areas, you cannot use the stand-alone tests to compare a student's relative strengths and weaknesses across curricular areas. For example, you could not say a student is better in reading than in mathematics.

Single-Level Standardized Tests

If you do not want to measure growth or development, a **single-level test** may be useful. Rather than cover several grade or age levels, such tests are directed toward one level or a particular course. Usually these assessments are built for high school and college courses. There are, for example, tests for Algebra I, first-year college chemistry, and first-year college French.

Each test is a stand-alone product and is not coordinated with other tests. Thus, these test results cannot be used to compare a student's relative standing in several subjects. Scores from this group of achievement tests are most often interpreted using norm-referencing schemes such as percentile ranks and standard scores.

If you are teaching in a single subject area, such as Algebra I or 19th-Century English Literature, you may be interested in assessing how well students are performing in just that subject. Multilevel tests are often inappropriate for such courses because they span several grades with relatively few items and thus lack content relevance for a particular course. *For most purposes, a teacher-made* *test for a subject is most appropriate: It is closest to the course content and contains the emphasis you desire.*

Single-subject or course tests have been found most useful for such purposes as pretesting to determine the general background of students coming into a course; advanced placement in college courses; exemption from required or introductory courses; contests and scholarship programs that reward general knowledge of a particular subject; and granting college credit for knowledge acquired by independent study, work experience, or other types of nontraditional education. Many tests for specific subjects are listed in the *Mental Measurements Yearbook*s.

STATE- OR DISTRICT-MANDATED TESTS AND CUSTOMIZED TESTS

State-Mandated Tests

States require students to sit for official **state-mandated assessments**, especially after the NCLB Act required testing all students in Grades 3–8 and high school. State-mandated tests vary greatly in their focus, makeup, and quality. Most state assessments have accountability as a focus. Accountability may be at the school district, school-building, or student level. The NCLB Act requires accountability at the school level in an attempt to ensure all students in the school receive quality instruction. At the district or school level there may be serious consequences if scores do not improve over time. In some states, schools failing to improve can be "taken over" by a team appointed by the state. In school-level accountability programs individual students may not receive their results. The test may be a high-stakes test for a school, but a low-stakes test for a student.

Some states require individual student accountability in addition to school accountability. This usually takes the form of a graduation test. The test may cover basic skills or be more challenging, depending on the state. Often a basic skills graduation test is given in Grade 9 or 10, so that students with low scores may be forewarned and placed into remedial programs to improve their skills. Graduation tests are high stakes for students.

State assessments are based on a state's curriculum framework and standards. The trend had been to make standards that are challenging to students rather than to limit them to minimum competencies or basic skills.

Customized state assessments are usually built and marked by a proprietary agency under a state contract. Test publishers tender bids in response to a state's request for proposals. Usually, the publisher winning the bid uses a secure form of the survey battery (it is parallel to the unsecured form) and then supplements the battery with additional test items that match state standards not covered by the original battery.

Customized state assessments are also built for states' alternate achievement standards, and sometimes for modified achievement standards. No Child Left Behind mandates that all students participate in state assessment. A small number of students with significant cognitive disabilities may be taught and assessed using expanded benchmarks (downward extensions of state standards) and alternate assessments. Some states are also experimenting with modified achievement standards and assessments for students whose cognitive disabilities are not so severe as to require the alternate achievement standards, but for whom the regular state standards and assessment may not be appropriate.

States have experimented with large-scale performance assessments and portfolios, with mixed success. The latter are costly to develop; time-consuming to administer; costly to score; and difficult to craft to high-technical standards of reliability, validity, and year-to-year comparability. Some states have persisted and done well, but many returned to either multiple-choice testing or a mixture of multiple-choice testing with a small portion of performance assessment and writing assessments.

As a teacher, you will no doubt find yourself working in a controversial state assessment environment. You will be required to implement and participate in your state's assessment program. Legislators may tend to blame teachers for what is wrong with the educational system. You may feel pressure for your students to do well on the test. You can read about state assessment programs and their controversies in a weekly periodical called *Education Week* (http://www.educationweek.com).

You can usually find out about your state's assessment program through its education department's Website. Websites for individual schools (if they exist) can be located by searching for the school name and location.

The consensus is that state assessment programs wanting to change classroom practices must place great emphasis on teacher development because assessment programs alone do not improve schools (Linn & Baker, 1997; McDonnell, 1997; Smith, 1997). In addition, compromises due to financial constraints, time pressures, technology limitations, and political pressures usually mean that the original plans for state assessment need to be scaled back. This may result in failing to implement key components such as improving classroom assessment practices, using performance assessments, or failing to implement appropriate teacher development programs.

Interim or Benchmark Assessments

Because of the high stakes attached to annual state testing, local districts have been interested in predicting and shaping students' performance during the year, to maximize performance on the annual state test. Test publishers have rapidly developed many products to meet this need. These tests can be called "benchmark," "diagnostic," "formative," and/or "predictive." Perie, Marion, and Gong (2007) call these tests by the umbrella term "interim assessment," a term which we think is appropriate, too, and which seems to be beginning to stick.

Following are some examples of products or suites of products marketed for interim or benchmark assessment. Some of these are administered online; others are paper-and-pencil tests or a combination. Some include item banks for creating customized interim assessments.

Acuity (CTB/McGraw-Hill)—Grades 3–8 and 10

Benchmark Tracker (Pearson)—Grades K–12

Measures of Academic Progress (*MAP,* Northwest Evaluation Association)—Grades 2–10

TerraNova Math and Reading Assessments (CTB/McGraw-Hill)—Grades K–12

Data Director (Riverside Publishing)—Grades K–12

Claims are made that interim assessments will help schools and districts meet adequate yearly progress requirements or improve performance on high school exit exams. A good interim assessment can be part of a district's balanced assessment system if the information is used well. At the present time, there is little research documenting prediction of state test scores or positive effects on student achievement (Brown & Coughlin, 2007).

Response to Intervention (RTI) Assessments

As Chapter 6 described, Response to Intervention refers to a process that emphasizes how well students respond to changes in instruction. The essential elements of an RTI approach are the provision of scientific, research-based instruction and interventions in general education; monitoring and measurement of student progress in response to the instruction and interventions; and use of these measures of student progress to shape instruction and make educational decisions (Klotz & Canter, 2006).

Most RTI solutions from test publishers include a variety of assessment and intervention tools and take a tiered approach that progresses from universal screening to progress monitoring and interventions. Examples include:

Yearly ProgressPro (CTB/McGraw-Hill)—Grades 1–8

Academic Intervention Monitoring System (AIMS, Pearson)—Grades K–12

Early Childhood Assessments

The impetus for early childhood assessments seems to be making certain that children are ready at younger age for the demands of accountability testing when they are older. Both academic readiness (Bracken, 2002) assessments for early childhood and interim assessments for pre-K to Grade 2 students have this general purpose. The *Bracken School Readiness Assessment (BSRA,* Pearson Assessments*)* is an academic readiness assessment for ages 3.0 to 6.11. The *Children's Progress Academic Assessment (CPAA,* Pearson*)* is an interim assessment system for pre-K to Grade 2.

English Language Proficiency Tests

English language proficiency testing is now a requirement under No Child Left Behind, Title III: Language Instruction for Limited English Proficient and Immigrant Students. The purpose of English language proficiency tests is to place English language learners (ELLs) at the appropriate level for bilingual or English as a second language programs, or determine if ELLs are ready to exit these programs. English language proficiency tests must test the domains of reading, writing, speaking, and listening, and must provide a comprehension score for English language learners from Grades K through 12. An additional requirement is that the tests measure progress from year to year. This is typically accomplished by placing the test levels on a developmental or vertical scale.

Examples of Enlgish language proficiency tests include:

LAS Links K-12 Assessments (CTB/McGraw-Hill)—Grades K–12

Stanford English Language Proficiency test (SELP, Pearson)—Grades Pre-K–12

ACCESS for ELLs (WIDA Consortium)—Grades Pre-K–12

NONSTANDARDIZED ACHIEVEMENT TESTS

Disadvantages

If a test is not standardized, its publisher probably has failed to try out the assessment materials extensively. It is likely that the publisher has not collected sufficient student-based data to support the quality of the test. There may be little or no empirical evidence of validity and reliability of the scores, and norms if they exist are unlikely to be based on a representative national sample.

Tests without empirical data to support them may be good tests, but because there is no documentation, you can't be certain the test's intended purposes are in fact being accomplished. As you read in Chapter 3, validity should be based on support from a variety of empirical data. Unfortunately, many school officials are ignorant of the principles of assessment validation and purchase such assessments, so these publishers stay in business while children suffer.

Textbook-Based Tests

Appeal Teacher's editions of texts often have assessment tasks at the end of chapters or at the back of the books. Some curriculum materials have assessment tasks built into the learning materials; others have separate tests for photocopying or duplicating. These assessment procedures have some appeal. If you follow the textbook closely, the tests' topics appear to closely match what you teach. And they are convenient. From a publisher's perspective, adding tests to curriculum materials is usually a marketing tool: Making the curriculum materials appealing to teachers increases the likelihood of adoption.

Disadvantages The quality of assessments that come with curriculum materials is usually poor. Although there may be exceptions, publishers rarely use trial data to improve these assessment materials. Trying out assessment materials for purposes of improvement is especially important when the assessment tasks are performance tasks, because wording strongly influences how students interpret the tasks, and scoring rubrics need to be refined using them with actual student responses. Also, a text-series author is seldom proficient in assessment development. The publisher's editorial staff does not edit the tasks for their technical assessment merits using checklists such as those found in Chapters 8 through 12 of this book. Sometimes the keyed answers to the questions are incorrect or contradict what is in the curriculum materials.

For example, Bob's parents helped him study for a social studies test. The following day he took it in class. A few days later he brought it home after the teacher marked it. When his parents reviewed it, they discovered two items were marked wrong although Bob gave the correct answer. The teacher's answer key contradicted what the book said. When this error was pointed out for the two items, the teacher insisted that such contradictions were impossible because the textbook publisher printed the test. This teacher was mistaken.

Often the questions on these test materials focus on low-level cognitive skills such as recalling facts and definitions. Despite these flaws, many teachers place unwarranted faith in these tests and use them to make important day-to-day decisions about students' learning progress.

What to Do What should you do when your curriculum materials contain appealing assessment tasks? First, don't accept their quality at face value, no matter how good the rest of the curriculum materials are. Look carefully at the embedded or "homework" type of exercises to be sure that they assess something worthwhile for students to learn, that the answers are correct, and that the tasks completely match your state's standards and your school district's important learning targets. Second, be prepared for disappointment: You may need to rewrite these exercises yourself.

Third, review carefully any tests or quizzes that come with the materials. You use the scores from tests and quizzes to determine students' grades, so you want these assessments to be of high quality. Review each item for correctness, importance, and match to standards and your learning targets. Use the checklists in Chapters 8 through 12 to help you edit the items, revising them when necessary. When you use the items in class, review your students' responses to them to help you discover flaws and correct them. Teachers can work together in small groups to review and improve assessment materials that come with curriculum materials. This can be done a little at a time, over 2 or 3 years.

APPROPRIATE USES OF STANDARDIZED TEST RESULTS

Within-Classroom Uses

How can you use standardized test results? Here are some suggestions for within-classroom uses of test results**:**

1. *Describe the educational developmental levels of each student.* Use this information about the differences among your students to modify or adapt teaching to accommodate individual students' needs.
2. *Describe specific qualitative strengths and weaknesses in students.* These strengths vary from one curriculum area to another. Use this information to remediate deficiencies and capitalize on strengths.
3. *Describe the extent to which a student has achieved the prerequisites needed to go on to new or advanced learning.* Combine these results with a student's classroom performance to make recommendations for placement.
4. *Describe commonalties among students.* Use this information to group students for more efficient instruction. Figure 15.5 outlines suggestions for using survey battery information in planning classroom instruction.

FIGURE 15.5 **A systematic procedure for using the results of a standardized achievement test to plan instruction for a class.**

Step 1. Review the class report to determine weaknesses

Use a report that summarizes performance on clusters of items for all students in your class. Within each curriculum area, identify on which clusters your students most need improvement. Match the clusters to your state's standards and determine the class's weakness and strengths with respect to the standards. Use your knowledge of the subject and of your students to verify the areas of greatest need. Don't be afraid to contradict the picture given by the test if you have good evidence that supports the fact that the students know more than they have shown on the test.

Step 2. Establish instructional priorities

Review your list of instructional needs. Put them into an order for instruction. Be sure to teach prerequisite needs first. Concentrate on the most important areas—those that will help students in their further understanding of concepts and principles in the subject.

Step 3. Organize the class for instruction

The test information may help you form small groups of students who have similar instructional needs. Alternately, you could form small groups that have students at different levels of learning so that those who already know the material can help instruct those who have not yet mastered it. You will need to use your own resources to organize your class, as the test cannot do that directly.

Step 4. Plan your instruction before you begin

Be clear about your instructional targets. Look at the test items to get an idea of the types of tasks you want students to learn to do, but remember that you are trying to teach generalizable skills and abilities. The tasks on the test are only a small sample of the domain of tasks implied by the curriculum.

Look to the curriculum to see where the areas of need fit into the larger scheme. Teach within this larger framework, rather than narrowing your teaching to the test items. Create your own assessment instrument for each of the areas of need so you can clarify what you will expect students to do at the end of the lessons. Organize your teaching activities to accomplish these ends.

Step 5. Assess students' progress toward your instructional targets and state's standards

Monitor students' progress through both informal and formal assessments. Observe students as they complete the assignments you give them to see if they are making progress toward your learning targets and state standards. Use performance and paper-and-pencil assessments to monitor their progress in more formal ways. Adjust your teaching for those students who are not making appropriate progress. Give feedback to students by showing them what they are expected to do (i.e., the learning target or state standard), explaining to them what their performance is like now, how it is different from the target performance, and what they have yet to learn to accomplish the target performance.

Step 6. Carry out summative assessment

Use a variety of assessment techniques to assess each student so that you are certain that the student has learned the target and can apply the concepts and principles to appropriate realistic situations. Use performance assessments, extended responses, and objective items in appropriate combinations. Do not limit your assessment to only one format.

5. *Describe students' achievement of specific learning targets.* Use students' performance on clusters of items to make immediate teaching changes.
6. *Provide students and parents with feedback about students' progress toward learning goals.* Use this information to establish a plan for home and school to work together.

Survey tests measure broad, long-term educational goals rather than immediate learning outcomes. It may take all year for a student to learn to read well enough, for example, to show some sign of improvement on a survey test. Meanwhile, however, the student may learn many specific skills and reading strategies. The student may perform well on your classroom assessments of these immediate learning targets.

Norm-referenced survey information is not likely to give you the fine-grained details you need to design an individual student's daily or weekly instructional plans. Classroom assessment procedures provide information about a student's performance in more specific areas. They are likely to be more useful to you for daily or weekly instructional planning than are ordinary survey tests. The results of survey tests can be used, however, to help you plan for a year or a term.

Standardized tests are often administered in the fall, after you have organized the class, and the answer sheets are sent away for scoring. By the time you receive students' results, several weeks of schooling have already passed. Such circumstances work against the possibility of using standardized tests for immediate instructional decisions. This is not to say, however, that you should disregard the results. Results from tests administered last spring will help you plan your teaching this fall.

Another use of standardized test results is to confirm or corroborate your judgment about a student's general educational development. It is important to realize that no single source of information about a student is entirely valid—be that source your own observations, results from assessments you developed, or results on standardized tests. Nevertheless, standardized tests can provide additional information that may alert you to the need to consider a particular student further.

Extraclassroom Uses

Standardized survey tests are also useful for extraclassroom purposes. Among these external uses of test results are the following:

1. *The average scores of a group (class, building, or school system) help school officials make decisions about needed curriculum or instructional changes.* The results provide one important piece of information if school officials judge the tests to be relevant and important to the goals of the local community.
2. *Test results also help school superintendents describe to parents, school boards, and other stakeholders the relative effectiveness of the local educational enterprise.* However, school board members should realize that no single instrument can account for all the factors that affect the learning of students in a particular community.
3. *Results help educational evaluators compare the relative effectiveness of alternate methods of instruction and describe some of the factors mitigating their effectiveness.*
4. *Results help educational researchers describe the relative effectiveness of innovations or experiments in education.*

INAPPROPRIATE USES OF STANDARDIZED TEST RESULTS

Criticisms of Standardized Tests

Criticisms of norm-referenced standardized achievement tests are quite common. Some critics find the very idea of a commercial, external, norm-referenced, summative, and/or quantitative device for measuring educational outcomes repugnant. Of course, any assessment procedure—standardized or not, norm-referenced or criterion-referenced, formative or summative, external or teacher-made, qualitative and quantitative—can be misused. Much of this misuse, moreover, comes not from something inherent in a particular standardized test, but from the invalid claims that persons make for some assessments or the unscrupulous way(s) in which an assessment might be used. Throughout this text, we discuss appropriate uses and emphasize the need to validate claims made for assessments. Use professional judgment in administering and interpreting all assessment procedures. The *Code of Professional Responsibilities in Educational Measurement* (Appendix C) and the *Code of Fair Testing Practices in Education (Revised)* (Appendix B) describe your responsibilities with regard to standardized achievement tests.

Criticisms of standardized tests may focus on some intrinsic characteristic of a test, such as its content coverage; something that is not part of a test, such as its failure to test certain student characteristics; or the misuse of test results, such as inappropriately using a test to classify or label a student. Some criticize several of these aspects of standardized testing. For example, they may say that tests (a) measure only a small portion of what is taught in the classroom (intrinsic characteristic), (b) do not measure the real goals of an educational program (characteristics not measured), and (c) foster undesirable changes in school curricula or teacher emphasis (misuse of test results).

Some criticisms are contradictory, and many of the criticisms can be overcome. The same test may be criticized by some persons because its focus is too narrow and by others because its scores are influenced by too broad a range of human characteristics. You may overcome many problems by either using the test in the way the publisher intended it to be used or by choosing another, more appropriate test.

Misuses of Standardized Tests

You should always strive to use the results of achievement assessments—survey batteries, performance assessments, or authentic tasks—in valid, professional ways. Never use a single assessment result to make an important decision about a student. Inappropriate uses of a survey achievement battery are listed here. Using a state-mandated assessment in these ways is also inappropriate.

1. *Placing a student in a special instructional program solely on results from a standardized achievement test.* Special programs include remedial programs as well as programs for students who are gifted and talented. School officials can overcome this misuse by using many pieces of information when making these decisions. They should include students' daily classroom performance, teachers' assessments, and results from other assessments in addition to the survey achievement battery.

2. *Retaining a student in a grade solely on the results from a standardized test.* First, you should recognize that the wisdom of retaining students is very much an open educational question, and the common practice of retention in the early grades often does not help students (Karweit & Wasik, 1992). Second, your daily observation, teaching, and evaluation of students are the most relevant types of information that a school official should use when making this decision. Third, parents have information about a child that school officials and teachers do not. Although standardized achievement test scores may have some bearing on this type of decision, their importance should have little weight in the final decision.

3. *Judging an entire school program's quality solely on the basis of the results from a standardized achievement test.* School programs are complex. They teach many things other than those assessed by standardized tests. You know, too, that even within a curriculum area assessed by a test, there is no perfect match between what is assessed and all the instructional targets in the curriculum framework. School officials can overcome misuse by aggressively placing program evaluation decisions in a broader context of the full curriculum framework and the full context of school and community factors.

4. *Using a survey achievement battery to prescribe the specific content teachers should teach at certain grade levels.* You know that a test only samples the many tasks that students could be asked to perform. Although test tasks are important, each task is not an end in itself. If the tasks are a representative sample from this larger domain, they allow you to generalize beyond them to estimate a student's performance on the domain. If school officials manipulate the sample or try to limit the curriculum domain primarily to the sample appearing on a test, they destroy the ability to generalize. School officials may overcome this misuse by developing curriculum frameworks using appropriate principles drawn from educational development, child development, learning, and the subject-matter disciplines. They should then select the test that best matches the important curriculum learning targets, rather than vice versa.

5. *Attributing a student's poor assessment results to only one cause.* Sometimes a teacher or school administrator interprets a student's assessment result as though it was entirely the result of the student's own shortcomings rather than the result of several interacting conditions. A student's poor assessment result may very well reflect the quality of previous teaching, the nature of the student's home environment, or other personal experiences.

6. *School officials or parents trying to blame the teacher if the class does poorly on a standardized test.* Before a person can attribute the rise (or fall) of a

class's test scores solely to a particular teacher, that person would have to consider how each of several factors influences the scores: Did the content of the test match the breadth and emphasis of what was taught in the classroom? Did the students in this year's class have, on the whole, better or worse general school aptitude than classes in the past (or classes assigned to other teachers)? Were students in this (or another) class taught the answers to the items or otherwise given an unfair advantage? Did last year's teacher do an exceptionally good (or poor) job of teaching, and did this influence carry over to this year? What home factors influenced the students' successes (or lack thereof)? Did the school principal (or other instructional leader) facilitate or inhibit the teacher's teaching or the students' learning? You can probably name other factors to consider when trying to find the reasons for a class's test results.

HOW TO ADMINISTER STANDARDIZED TESTS

You will most likely be required to administer one or more standardized assessments per year. These may be standardized achievement tests, performance assessments, or assessments mandated by your state department of education. Part of the validity of your students' results will depend on how well you follow the standardization procedure specified in the teacher's administration manual.

The Right Way to Prepare Yourself and Students for a Standardized Test

There are two important areas of assessment administration that you directly control and that directly affect the validity of your students' results. One area is how you prepare yourself and the students for the assessment. Standardized assessments, regardless of whether they are performance or multiple-choice formats, require students to be aware of (a) the fact that they will be assessed, (b) what they will be assessed on, (c) the reasons for the assessment, and (d) how their results will be used. Students should be prepared to do their best. You must also be prepared to administer, and perhaps to mark, the assessments. That means you must be familiar with the assessment procedures and materials, prepare the assessment environment so that a valid assessment can be done, understand how to administer the assessment—including what you are permitted to say to the students—and know how to prepare the students for the assessment.

The Right Way to Administer a Standardized Test

A second area in which you need to perform well is in actually administering the assessment. Valid assessment results will depend on how well you carry out your responsibilities during the administration phase. You need to follow the procedures stated in the manual exactly: Otherwise, the assessment results will not be comparable across students and using the norms will be invalid. Also, you need to monitor students to be sure they are following directions, marking their answers in the proper manner, and otherwise attending to the tasks. Figure 15.6 is a checklist of what you must do to administer a standardized assessment without lowering its validity.

ETHICAL AND UNETHICAL STUDENT PRACTICE FOR STANDARDIZED TESTS

The question of what type of practice to give students before they take a standardized assessment is an important one for you to answer. Educators do not agree about what is appropriate (Cohen & Hyman, 1991; Mehrens, 1991; Mehrens & Kaminski, 1989; Popham, 1991). The controversy concerns ethical test preparation practices. If you prepare students in inappropriate ways, then the validity of their assessment results is questionable. Do certain preassessment activities give your students unwarranted advantages that are not available to other students? If your students receive certain types of practice, can you or others still validly interpret their scores? If you teach your students certain responses or answers, can you generalize their assessment results properly?

A Clearly Unethical Teaching Practice

One of the guiding principles for ensuring validity is the **generalizability of assessment results**. (See Figure 3.2.) That is, can you infer a student's performance on the entire curriculum domain from the specific items the student took? If not, the validity of the results is low. For example, suppose there are 100 key concepts in a particular area of social studies. Further, suppose that instead of teaching students strategies for organizing and understanding these concepts and principles, you

FIGURE 15.6 Checklist for administering a standardized assessment procedure to your students.

Before the assessment date

1. Prepare a schedule for assessment, including dates and times for each component you need to administer.
2. Discuss the upcoming assessments with the students.
   a. Explain the purpose of the assessment.
   b. Explain what they will be doing.
   c. Explain when and how they will receive the results.
   d. Explain how the results will be used.
3. Become familiar with the assessment procedure and the directions for administering it. (Practice taking the assessment yourself.)
4. If proctors are necessary, schedule and train them.
5. Be sure that all the assessment materials are available, that students have pencils and other necessary tools, and that scratch paper and other materials are available.
6. Make any necessary physical adjustments to the room.
7. Make a sign that reads, "Assessing. Please do not disturb us!" Use the sign during the assessment sessions.

During the assessment

1. Follow the directions exactly as given in the directions manual.
2. Monitor students to be sure they are working on the correct pages and activities and are recording their responses properly.
3. Supervise the work of any proctors that are present.
4. Make notes describing any irregularities, either for individual students or for the entire group.

Source: H. D. Hoover, A. N. Hieronymus, D. A. Frisbie, and S. B. Dunbar, 1993. Copyright © 1993 by The University of Iowa. All rights reserved. Adapted from the *Iowa Tests of Basic Skills, Directions for Administration, Forms K and L., Levels 9–14*, p. 15, with permission of the Riverside Publishing Company.

picked only the four concepts that will appear on a standardized social studies test and taught answers only to the questions about those four concepts. Assuming you are a good teacher, your students would do very well on the test questions related to these concepts, and their test scores would be higher. However, your students would most likely not understand or integrate the broader social studies framework and the full set of concepts the course was supposed to teach. In other words, by narrowing your teaching to only those few tasks that appear on a specific test, you have failed to provide your students with empowering strategies to organize social studies concepts and principles. Further, you cannot interpret their test results as reflecting their general knowledge of the course concepts and principles. By teaching only those four concepts, you invalidated the students' test results and corrupted the students' education.

When you assess students, you want to generalize from their performance to the larger and broader domain of abilities and knowledge that the curriculum framework is supposed to foster. Responses on a particular test or assessment are only signs or pointers to the students' possible performance in the larger domain implied by the learning targets of the curriculum framework. However, if you give specific practice only on the questions or tasks on the assessment, you focus students' learning only on these few tasks. It is very unlikely that such narrowly focused instruction and learning can generalize to the broader learning targets that are the real goals of education.

The Range of Ethical to Unethical Practices

You can provide a variety of practice activities to help students improve their performance on an assessment. Which of these is appropriate? The following list of assessment preparation activities is arranged in order from the most to the least legitimate (Haladyna, Nolen, & Haas, 1991; Mehrens & Kaminski, 1989):

1. Teaching the learning targets in the curriculum without narrowing your teaching to those targets that appear on a standardized assessment.
2. Teaching general test-taking strategies, such as those discussed in Chapter 13.
3. Teaching only those learning targets that specifically match the targets that will appear on the standardized assessment your students will take.
4. Teaching only those learning targets that specifically match the targets that will appear on the

standardized assessment your students will take and giving practice using the same types of task formats that will appear on the assessment.

5. Giving your students practice on a published parallel form of the assessment they will take.
6. Giving your students practice on the same questions and tasks that they will take later.

Most educators would agree that the first activity is always ethical because it is the teacher's job to teach the official curriculum. Most educators would also agree that the second activity, teaching students how to take tests and do their best on them, is not unethical. The fifth and sixth activities would always be considered unethical because they narrow instruction to only the specific assessment tasks that your students will be administered and practically eliminate your ability to generalize from the assessment results to the performance domain specified by the curriculum.

Thus the boundary between ethical and unethical test preparation practices falls somewhere between Activities 3 and 5. They indicate that the deciding factor lies in the degree to which a school wishes to generalize the test results. The closer the activity is to the fifth one, the less able are school officials to generalize students' assessment results to the official curriculum—unless, of course, the official curriculum is identical to the assessment instrument.

Koretz and Hamilton (2006) summarize a somewhat broader set of test preparation steps that have been documented as responses to recent high-stakes testing: teaching more, working harder, working more effectively, reallocation, alignment, coaching, and cheating. Their criterion for positive or desirable preparation is the generalizability of test results, and thus test preparation is desirable if it produces "unambiguously meaningful increases in scores" (p. 548).

Teaching more, working harder, and working more effectively all therefore constitute positive test preparation, because increases in scores would mean increases in learning in the whole domain (reading, mathematics, etc.). Reallocation of instructional time and resources, coaching, and other practices, sometimes done in the name of "alignment," that narrow the domain to only what is covered by the sample of test items are negative consequences of high-stakes testing, because increases in scores would mean increases only in the sampled part of the domain. And Koretz and Hamilton's last category, cheating, never produces a valid increase in scores.

CONCLUSION

This chapter has explored the meaning of the term *standardized test*. What are standardized are the conditions of administration, procedures, and scoring, so that as far as possible scores are comparable across time and place. The chapter discussed the most common kinds of standardized achievement tests, along with their purposes and uses, and described how to follow directions for test administration. Finally, the chapter discussed ways to prepare students for standardized tests. The next chapter turns to interpreting the various kinds of norm-referenced scores that are provided in standardized test results.

EXERCISES

1. Using test publishers' catalogs, the *Mental Measurements Yearbook*s, and other resources, identify one published test that fits into each category of the authors' scheme for classifying published achievement tests. Share your findings with your classmates.
2. Describe the students, their community, and subject(s) that you teach (or plan to teach). Through self-reflection, give specific examples of how you may misuse achievement test results in this context in each of the following ways. Share your findings with the others in your course.
   a. Failing to consider measurement error when interpreting a student's scores.
   b. Using only the test results for making a decision about a student.
   c. Uncritically interpreting a student's score as measuring a pure trait.
   d. Failing to consider the complex nature of the causes for a particular student's test performance.
3. Evaluate the appropriateness of each of the following standardized test preparation practices.
   a. The school uses the latest version of a certain test. A teacher uses a version of the test that is no longer being administered in the school to give students special practice.

b. A teacher copies items from a test that is currently being used in the school and gives these to students for practice.
c. A teacher teaches students general rules and strategies for taking standardized tests, such as how to eliminate options and "guess" when they are not certain, and how to plan their testing time wisely.
d. The curriculum framework calls for learning the grammar rules covered by the test the school uses. The teacher teaches the students how to use these rules to answer the same format of questions that will appear on the test, but does not provide practice in more natural contexts of writing sentences and paragraphs.
e. The curriculum framework calls for learning the grammar rules covered by the test the school uses. The teacher teaches the students how to use these rules to answer the same format of question that will appear on the test, but also teaches them how to apply the rules in their own writing of sentences and paragraphs.
f. A deaf student who is mainstreamed in an inclusive program plans to go to a special postsecondary school for deaf students. For admission, the postsecondary school requires the student to submit results from standardized reading and mathematics tests. The teacher gives the upcoming tests to the student to take home to read a few days ahead of time, then answers any clarifying questions the student has about the vocabulary and the type of strategies that should be used when answering the questions. Later in the week the teacher administers the tests to the student under standardized conditions but with the help of a sign language interpreter.

4. Using the Internet, locate three states' education departments and descriptions of their state assessment program. (A comprehensive list of state Websites is found on the U.S. Department of Education's Website: http://www.ed.gov/about/contacts/state/index.html?src=gu.) If your state has an assessment program, be sure to include it as one of the three. Compare the assessment programs in terms of student versus school accountability; objective versus constructed-response assessment; use of standards, teacher development, and capacity building; and general objectives and purposes. Share your findings with others in this course.
5. Figure 15.7 lists various types of tests across the top and various characteristics as row headings. For each characteristic, describe the extent to which it is found in each type of test. In the cells in the body of the table, mark:
   a. ++ if most tests in that category exhibit this characteristic.
   b. + if a few tests in that category exhibit this characteristic.
   c. 0 if it is very rare that tests in that category exhibit this characteristic.

FIGURE 15.7 **Comparisons of the characteristics of various kinds of published tests with teacher-made tests.**

		Published tests in the marketplace								
		Standardized, empirically documented				Nonstandardized, not empirically documented				Teacher-made tests
	Characteristic	Survey batteries	Criterion-referenced tests	Other single-area tests	State-mandated accountability test	Single-course tests	Interim assessments	Criterion-referenced tests	Textbook accompaniments	
Content/ objectives covered	**1.** Common to many schools									
	2. Specific to one teacher/ school									
	3. Specific to one text or set of materials									
Intended to measure	**1.** Growth over time									
	2. Status on each specific objective in domain									
	3. Profile of strengths and weaknesses									
Norm-referencing provided	**1.** Several types of scores									
	2. Several types of norm groups									
	3. Spans several grades									
Criterion-referencing provided	**1.** Many items per objective									
	2. Diagnosis possible									
Provides materials for interpreting scores to	**1.** Students									
	2. Parents									
	3. Teachers									
	4. Administrators									
Technical quality	**1.** Professionally written items.									
	2. Empirical data on reliability and validity									

Interpreting Norm-Referenced Scores

KEY CONCEPTS

1. A referencing framework is a structure used to compare a student's performance to something external to the assessment, in order to interpret performance. A norm-referencing framework interprets a student's assessment performance by comparing it to the performance of a well-defined group of other students who have taken the same assessment. A criterion-referencing framework interprets a student's performance according to the kinds of performances a student can do in a domain. A standards-referenced framework combines elements of both.
2. Use normative information to describe student strengths, weaknesses, and progress.
3. Test publishers may provide norm-referenced scores based on information from several different norm groups.
4. Different types of norm-referenced scores are constructed to serve different purposes.
5. The percentile rank tells the percentage of the students in a norm group who have scored *lower* than the raw score in question.
6. A linear standard score tells how far a raw score is from the mean of the norm group, expressing the distance in standard deviation units.
7. A normal distribution is a mathematical model (an equation) based on the mean and standard deviation of a set of scores.
8. Normalized standard scores are based on transforming raw scores on an assessment to make them fit a normal distribution.
9. Developmental and educational growth scales are norm-referenced scores that can be used to chart educational development or progress.
10. An extended normalized standard score tells the location of a raw score on a scale that is anchored to a lower grade reference group.
11. A grade-equivalent score tells the grade level at which a raw score is average.
12. Five guidelines for score interpretation will serve you well: look for patterns in scores, seek explanations for the patterns, don't expect many surprises, don't overinterpret small differences, and use evidence from other assessments to clarify interpretations.

IMPORTANT TERMS

area under the normal curve
derived scores
empirical norming dates
extended normalized standard score
grade-equivalent scores (*GE*)
grade mean equivalent
interpolation versus extrapolation
IRT pattern scoring
item response theory (IRT)
linear standard scores(*z*,*SS*)

From Chapter 16 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

modal-age norms

normal curve equivalent (*NCE*)

normal distribution

normal growth (grade-equivalent view, percentile rank view)

normalized standard scores (z_n, *T, DIQ, NCE, SAT*)

normalizing a set of scores

norm groups (local, national, special)

norm-referencing versus criterion-referencing

percentile ranks (local and national)

raw scores

relevance, representativeness, and recency of norm data

SAT-score

school averages norms

standards-referencing framework

stanine scores (national stanine)

THREE REFERENCING FRAMEWORKS

Suppose that you took a spelling test and your score was 45, found by giving one point for each correctly spelled word. How well have you performed? Knowing only that your task was "a spelling test" and that your score was 45 leaves you unable to interpret your performance.

Raw scores are the number of points (marks) you assign to a student's performance on an assessment. You may obtain these marks by adding the number of correct answers, the ratings for each task, or the number of points awarded to separate parts of the assessment. As in the preceding spelling score example, a raw score tells a student what he or she "got," but says very little about the *meaning of the score*.

Practically all educational and psychological assessments require you to use some type of referencing framework to interpret students' performance. A *referencing framework* is a structure you use to compare a student's performance to something external to the assessment itself. In Chapter 14, we discussed referencing frameworks in the context of grading.

Norm-Referencing Framework

Norm-Referencing A **norm-referencing framework** interprets a student's assessment performance by comparing it to the performance of a well-defined group of other students who have taken the same assessment. The well-defined group of other students is called the **norm group**. To make valid norm-referenced interpretations, all persons in the norm group must have been given the same assessment as your students under the same conditions (same time limits, directions, equipment and materials, etc.). This is why you must follow administration instructions exactly when administering a standardized achievement test whose results you later will want to interpret through a norm-referenced framework.

To understand a norm-referenced interpretation, let's return to your score on the spelling test. Suppose your raw score of 45 means that your percentile rank (*PR*) is 99—that is, 99% of the persons who took the spelling test have scored lower than 45. Before you congratulate yourself, however, you should determine who is in the norm group to which your raw score is being referenced. You would interpret your performance differently if you knew the norm group was composed of third graders than if the norm group comprised adults.

Validity of Norm-Referenced Interpretations Your norm-referenced interpretations are less valid when the norm group is not well defined. The more you know about who is in the norm group, the better you can interpret a student's performance in a norm-referenced framework. Consider the difference in interpreting your performance on the spelling test, for example, when the norm group is adults in general versus a norm group composed of adults who have won prizes in national spelling contests.

Norm-Referenced Scores **Derived scores** make norm-referenced interpretations easier. A more or less standard set of derived scores is now routinely reported for most published tests in education:

1. *Percentile ranks* tell the percentage of persons in a norm group scoring lower than a particular raw score.
2. *Linear* ***standard scores*** tell the location of a particular raw score in relation to the mean and standard deviation of a norm group.

3. *Normalized* ***standard scores*** tell the location of a particular raw score in relation to a normal distribution fitted to a norm group.
4. *Grade-equivalent scores* tell the grade placement for which a particular raw score is the average for a norm group.

Criterion-Referencing Framework

Beyond Norm-Referencing Norm-referencing is not enough to interpret your score fully: You may be a better speller than other people—whoever they happen to be—but what can you spell? At a minimum, you would need to know the kinds of words in the pool from which those on the spelling test were selected, the number of words selected, and the process used to select the words. Were they really words, or were they nonsense syllables? Were they English words? Were they selected from a list of the most difficult (or easiest) English words? Did the test have 45 words or 500 words? Did the words on the test represent some larger class or domain? Did spelling the words require you to use certain mental processes or to apply certain spelling rules?

These questions are especially important when you need to make absolute interpretations of students' assessment performance—for example, when you need to know which specific learning target your students are having trouble mastering. Norm-referencing provides information to help in your relative interpretations of scores, but frequently these are not enough. Scores that reflect relative achievement such as rank order, for example, may be helpful in picking the best readers, or in sectioning a class into better, good, and poor readers. However, to plan appropriate instruction, eventually you need to know each student's specific reading skills and the particular types of difficulties each student is experiencing. When your diagnosis and prescription are based on students' error patterns or on your analysis of their faulty reasoning or thinking processes, as described in Chapter 7, you must put aside the norm-referencing framework and use a criterion-referencing framework.

Criterion-Referencing You use a **criterion-referencing framework** to infer the kinds of performances a student can do in a domain, rather than the student's relative standing in a norm group. This domain of performance to which you reference a student's assessment results is called the *criterion*. When you teach, the criterion that is of most interest is the domain of performance implied by your state's standards, your curriculum framework, and your lessons' learning targets.

Validity of Criterion-Referenced Interpretations

Your criterion-referenced assessment interpretations lose validity when the domain of performances to which you wish to infer your students' status is poorly defined, or when your assessment is a poor sample from that domain. The more you know about the domain from which the tasks on your assessment were sampled, the more validly you can interpret their results. For example, if you did not construct your assessment using clearly defined statements of learning targets, or if your assessment inadequately represents the wide range of performance implied by a clearly defined set of learning targets, then you have only a weak basis for making criterion-referenced interpretations.

You can easily see why by reviewing the spelling example again. Suppose you knew that the spelling domain was the 10,000 most frequently misspelled English words, and that the assessment had been constructed as a sample of 100 words representative of the spelling patterns in this domain. In this case you may interpret your score of 45 on a 100-word assessment as an estimate of the proportion of those 10,000 words you know how to spell. You can see that if there were only 50 words on the assessment, your estimate would be less accurate than when there are 100. A sample of 10 words is even less accurate. Further, if the 100 words did not sample the domain representatively, your estimate also would be less accurate, even though there were 100 words. For example, the 100 words may contain only regular spelling patterns and ignore others. Thus both the number of items on the assessment and how well they represent the domain contribute to how valid your criterion-referenced interpretation is.

Criterion-Referenced Scores Criterion-referenced assessments do not have well-developed, derived score systems like norm-referenced assessments. Nevertheless, certain types of scores are often used with these assessments:

1. *Percentage*—a number telling the proportion of the maximum points earned by the student (percentage correct, percentage of objectives mastered, etc.).

2. *Speed of performance*—the time a student takes to complete a task, or the number of tasks completed in a fixed amount of time (typing 40 words per minute, running a mile in 5 minutes, completing 25 number facts correctly in 1 minute, etc.).
3. *Quality ratings*—the quality level at which a student performs ("Excellent," rating of "5," "mastery," etc.).
4. *Precision of performance*—the degree of accuracy with which a student completes a task (measuring accurately to the nearest 10th of a meter, weighing accurately to the nearest gram, fewer than 10 typing errors, etc.).

Standards-Referencing Framework

Meeting Standards The NCLB Act of 2001 requires states to report the percentage of students who have achieved at three levels—basic, proficient, and advanced—in meeting a state's reading, language arts, mathematics, and science standards. The three levels of achievement in each subject area are specific to a state's particular standards. States use tests that are aligned with their standards to classify a student as attaining basic, proficient, and advanced achievement in each subject area. Because the goal of the NCLB Act is to have all students achieve at the proficient or higher levels (as these levels are defined in each state), there is an additional federal mandate that states show that they are making adequate yearly progress toward achieving this goal.

Standards-Referencing The NCLB's accountability standards require that students' scores on a test be referenced to the standards-defined achievement levels. The immediate testing question that a state faces is deciding what range of test scores is to be called "basic," what range is "proficient," and what range is "advanced." Once these ranges of scores are defined, students' scores are referenced to those ranges and interpreted to mean basic, proficient, or advanced achievement in a subject. This is called a **standards-referencing framework** (Young & Zucker, 2004).

Combining Frameworks The standards-referencing framework is accomplished by combining aspects of the criterion-referencing and the norm-referencing frameworks. On the criterion-referencing side, test items are selected to match or align with the state's standards. On the norm-referencing side, the state administers the test to the students and gathers information about the performance of students on each test item. A common procedure is then to order the items from easiest to most difficult. Panels of experts (including teachers) use this ordered list of test items, along with their knowledge of the subject area and students, to set the score that forms the boundary between each achievement level. If a student's score falls between the lower and upper boundaries of a category, the proficiency category, for example, then the student is classified into that category (e.g., proficient.)

Adequate Yearly Progress The boundaries for basic, proficient, and advanced are set using the population of students who took the test the first year. This group serves as the baseline group, so that in subsequent years, the state can measure whether its yearly progress is adequate. It does this by determining each year the percentage of students who have scores within each achievement level. Statistical and practical rules are established to determine whether the percentage of students in the proficient and advanced categories increase enough to represent adequate yearly progress. The goal, as we said, is for all students in a state to have scores within or above the proficient level.

USING NORMS

Importance of Norms

Norm-referencing indicates how one student's performance compares to the performances of others. However, simply comparing students with one another is not a very good reason for assessing them (Hoover et al., 1993b). Here are the major reasons for assessing students:

1. To describe, within each subject area, the performances a student has achieved.
2. To describe, within each subject area, student deficiencies that need further improvement.
3. To describe, across the curriculum, which subjects are the student's strengths and weaknesses.
4. To describe, within each subject area, the amount of educational development (progress) a student has made over the course of one or more years.

The first two purposes are best served within a criterion-referencing framework. In essence, this requires you to look carefully at a student's performance, item by item, and compare it to your learning targets.

The second two purposes are best served within a norm-referencing framework. A student's relative strength in reading and mathematics, for example, cannot be described on purely a criterion-referenced basis. You can describe what a student can do in each area, but you need a norm basis to conclude whether these are relative strengths or weaknesses. A teacher may say, for example, that a student is able to solve routine linear and quadratic equations in mathematics and is able to read with comprehension age-appropriate stories. However, which is the stronger area? Normative information can determine this.

Standardized tests describe students' relative strengths and weaknesses in different curricular areas because of the normative information they provide. The same group of students at the same grade level (the norm group) is administered tests covering several curricular areas. Thus, if fourth grader Blake ranks at the top of the norm group in mathematics but in the middle of the norm group in reading, we know that of the two subjects, Blake is stronger in mathematics.

The fourth purpose mentioned earlier—measuring educational growth and development—also requires norm-referencing. Norm groups provide the basis for defining an educational development scale (such as the grade-equivalent scale) across different grade levels. We assess a student once every year or two, each time referencing the results to this developmental scale. We measure growth by the student's progress along this scale.

The remainder of this chapter discusses the various norm-referenced scores and scales used in educational assessment. As a teacher, you will not be required to create growth scales or calculate scale scores. However, you will be required to interpret and to use such scales and scores with your students. In addition, you will be expected to explain the meaning of reports of these scores to your students and their parents.

TYPES OF NORM GROUPS

Before you can understand and use norm-referenced scores you need to understand the meaning of norms and norm groups. As we've already stated, a norm group is the large representative sample of students for which test manuals report performance. The performance of a norm group on a particular assessment represents the present, average status of that group of students on that particular assessment. A group's current average does not represent a standard, however, nor does it establish what your school or your students should attain. Your state's content and performance standards and your curriculum's learning targets tell you what students should achieve. Comparing your students and school to norm groups can help, however, decide the general range of performance to expect from your students, provided your students are similar to those in the norm group. As you will see, test publishers may provide information on several different groups when reporting norm-referenced scores.

Multiple Norm-Group Comparisons

Ordinarily, a student is a member of more than one group. For example, a 14-year-old, eighth-grade boy with a hearing impairment took a standardized mathematics concepts test and obtained a raw score of 32. This may represent a percentile rank of:

- 99 in a national group of hearing-impaired eighth graders.
- 94 in the test publisher's national eighth-grade standardization sample.
- 89 in the group of eighth graders in his local community.
- 80 in the group of eighth graders currently enrolled with him in an advanced mathematics course.

Depending on the decisions you must make, referencing a student's score to more than one norm group may be in order. Vocational counseling decisions, for example, may require that you compare a student's profile of abilities and achievements to each of several occupational or vocational groups about which the student is seeking career information. Comparing the person only to "students in general" may offer less information for career exploration.

Local Norms

For many of your norm-referenced interpretations, the most appropriate group with which you should compare a student is the **local norm group**: the group of students in the same grade in the same school district. It is this group with which you and the students will interact the most. Local percentile ranks or standard scores are easy to compile for a school's testing program, and your director of testing should provide them to you every time a standardized test is administered. Publishers also offer this service for their customers—frequently at extra cost, however.

National Norms

Most norm-referenced, standardized achievement and aptitude batteries have what are called **national norms**. In principle, the national norm groups are supposed to be representative of the students in the country, and some publishers expend a great deal of effort to ensure representativeness. But each publisher uses a somewhat different definition of what constitutes a truly representative national sample and conducts the sampling processes differently. The result is that the norms from different publishers are not comparable. You should note, however, that no publisher's norming sample exactly mirrors the nation's schools. A school's participation in a publisher's norming sample is voluntary. Sometimes this creates a self-selection bias in a given publisher's norms that may distort the norms in favor of schools that have used that publisher's tests in the past (Baglin, 1981). A more detailed description of how publishers obtain norming samples is given in Chapter 17.

National norms need not be composed simply of students in general at a grade level. A publisher may provide separate male/female norms or may provide separate norms for students with certain disabilities. Sometimes modal-age norms are provided. **Modal-age norms** include, from among all students at a particular grade level, only those near the most typical chronological age for that grade.

Special Norm Groups

For some tests **special norm groups** are formed. Examples include students with deafness or blindness, students with developmental disabilities, students enrolled in a certain course of study or curriculum, and students attending regional schools. A student may belong to more than one special group, of course.

School Averages Norms

School averages norms consist of a tabulation of the average (mean) score from each school building in a national sample of schools and provide information on the relative ordering of these averages (means). This distribution of averages is much less variable than the distribution of individual student scores. Figure 16.1 illustrates this difference in variability for one publisher's reading test.

If your school principal wants to know how the school's third-grade average score compares with that of other school buildings, then the principal needs to use school averages norms. For individual students' norms, a distribution of individual scores is made and used as the basis for norm-referencing. But individual student scores vary widely, so much so that comparing a school's average to that group may lead to misinterpretations.

Study the following example to get some idea of what your school district may gain from using school averages norms:

Example

Example of how using the wrong norms may lead to underestimating how well a school is doing

In Lincoln School the average spring fifth-grade developmental standard score on the *Iowa Tests of Basic Skills* (Reading Comprehension subtest) is 250 (see Figure 16.1). The principal looked up this number in the individual student norms table and erroneously concluded that the school ranks higher than 85% (*PR* = 85 for individuals) of the schools. (In Figure 16.1, look at the row labeled "*NPR* of Avg. *SS*: Student Norms.") Actually, the school is much better, ranking at the top 1% (*PR* = 99 for school averages, "*NPR* of Avg. *SS*: School Norms").

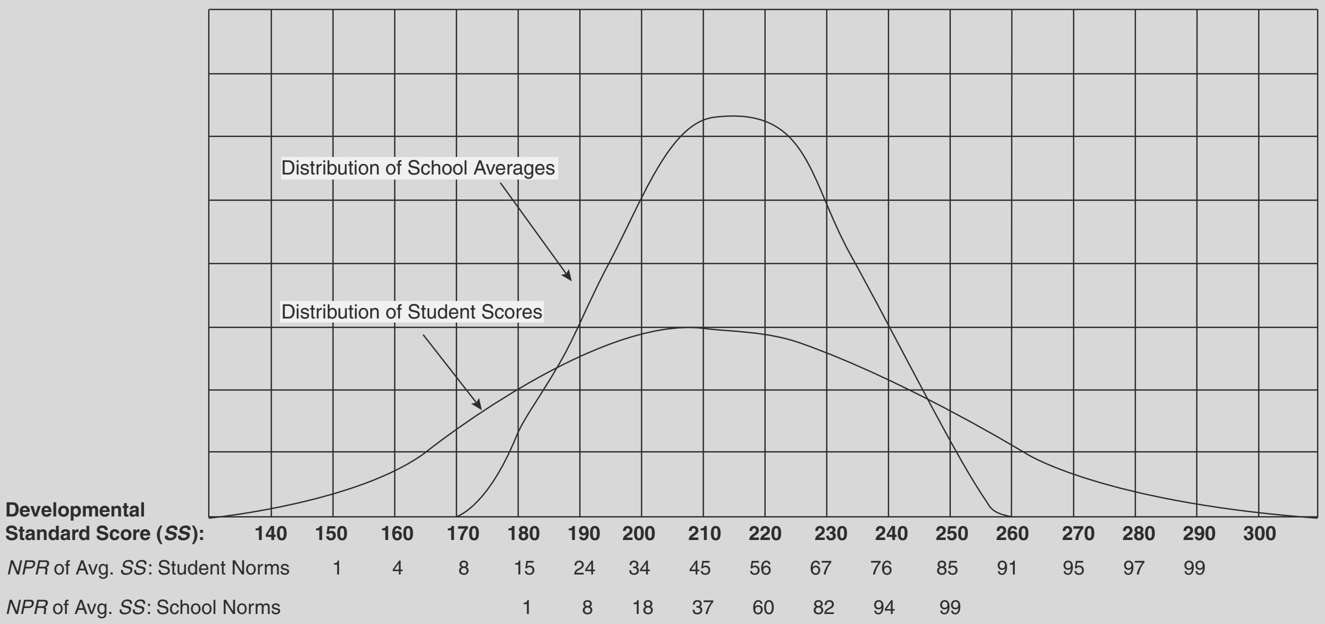
In general, if someone uses individual score norms erroneously and the school is above average, the results will underestimate that school's standing among other schools; those whose schools are below average will overestimate their standing among other schools. You can verify this principle by checking several developmental standard score values and percentile ranks in Figure 16.1.

Not all publishers provide school averages norms. Some publishers say that school averages norms mix together very small and very large schools. They say that mixing schools that are very different makes the data in a school averages norms table difficult to interpret correctly.

Using Publishers' Norms

Know When the Assessment Was Normed You obtain the most accurate estimate of a student's

FIGURE 16.1 Comparison of the distributions of students' scores and school averages for the Reading Comprehension subtest of the *Iowa Tests of Basic Skills*, Grade 5, spring norms.



Note: *NPR* = national percentile rank

Source: From S. B. Dunbar, H. D. Hoover, D. A. Frisbie, and K. R. Oberley, 2008. Copyright © 2008 by The University of Iowa. All rights reserved. *The Iowa Tests of Basic Skills, Interpretive Guide for School Administrators, Forms A, B, and C. Levels 5–14,* p. 79, with permission of the Riverside Publishing Company.

standing in a norm group when the student is tested on a date nearest the time of year the publisher established the norms. Publishers commonly interpolate and extrapolate to develop norm tables: They may provide spring norm tables, for example, even though no tests were actually administered to the norm group in the spring. Each publisher's empirical norming dates are different, but the publisher should state the dates in the test manual or technical report. To be accurate, your school should administer a standardized test within 2 or 3 weeks before or after the midpoint date of the publisher's empirical norming period.

Criteria for Evaluating Norms It is generally accepted (AERA et al., 1999) that published norms data should satisfy three Rs: relevance, representativeness, and recency. **Relevance** means that the norm group(s) a publisher provides should be the group(s) to which you will want to compare your students. **Representativeness** means that the norm sample must be based on a carefully planned sample. The test publisher should provide you with information about the subclassifications (gender, age, socioeconomic level, etc.) used to ensure representativeness. Remember that the sample size is not as crucial as its representativeness. Of course, if the population of students is very large, a representative sample should necessarily be large.

Recency means that the norms are based on current data. As the curriculum, schooling, and social and economic factors change, so too will students' performance on tests. Further, if your school uses the same form of a test year after year, scores will generally increase because the students become familiar with the format, and teachers tend to prepare students specifically for that test (Linn, Graue, & Sanders, 1990; Shepard, 1990; Wiser & Lenke, 1987). If the norms are not recent, they will mislead, conveying the impression that your students are learning better than they really are.

Using Norms Tables

Test manuals contain tables—called *norms tables*—for converting raw scores to different kinds of norm-referenced scores. No computation is required: You need "only" look up the score. Only is in quotes because looking up scores in a table properly and accurately is not as easy as it sounds. Specimen tables are shown later in this chapter, along with a discussion of the particular scores, so that you can practice using the tables.

NORM-REFERENCED SCORES

Norm-referenced scores are derived from the raw scores of an assessment. You should be aware that many types of norm-referenced scores exist. Space permits discussion of only the ones you will most often encounter, which are represented in the concept map shown in Figure 16.2.

Norm-referenced tests use, on average, more difficult items than classroom tests. This is in contrast to testing when the purpose is to describe students along a standards-based continuum of achievement (e.g., Basic, Proficient, Advanced; or A, B, C); then items should cover the range of difficulty levels to be described.

PERCENTILE RANKS

We begin at the leftmost branch of norm-referencing schemes in Figure 16.2. The **percentile rank** tells the percentage of the students in a norm group that have scored *lower* than the raw score in question. The percentile rank is perhaps the most useful and easily understood norm-referenced score. Figure 16.3 is an example of a publisher's norms table that gives percentile ranks for each raw score.

To read the norms table, locate the raw score obtained from the assessment in its correct column in the body of the table, and read out the corresponding percentile rank. For example, suppose a seventh grader named Veronica takes the *Differential Aptitude Tests (DAT)* on October 23, and she scores 48 in Mechanical Reasoning. Her percentile rank from the norms table in the last example is 98. She is above average in the norm group of seventh-grade females in mathematics; her raw score exceeds 98% of the females in the standardization group.

Notice there are three sets of percentile ranks in the example table—one for the seventh-grade boys, one for the seventh-grade girls, and one for the combined group. This is common practice for norm-referenced assessments in which there are large differences between males and females.

A raw score of 48 has a percentile rank of 90 for boys. This lower percentile rank for boys for the same raw score reflects that seventh-grade boys do much better as a group on this Mechanical Reasoning test. As a result, 48 does not rank as high for boys as it does for girls. When the boys and girls are combined, the resulting distribution is shown in the Combined column of the table in the last example.

Which gender norms should teachers and counselors use? The answer depends on how they

FIGURE 16.2 Organization of major score-referencing schemes.

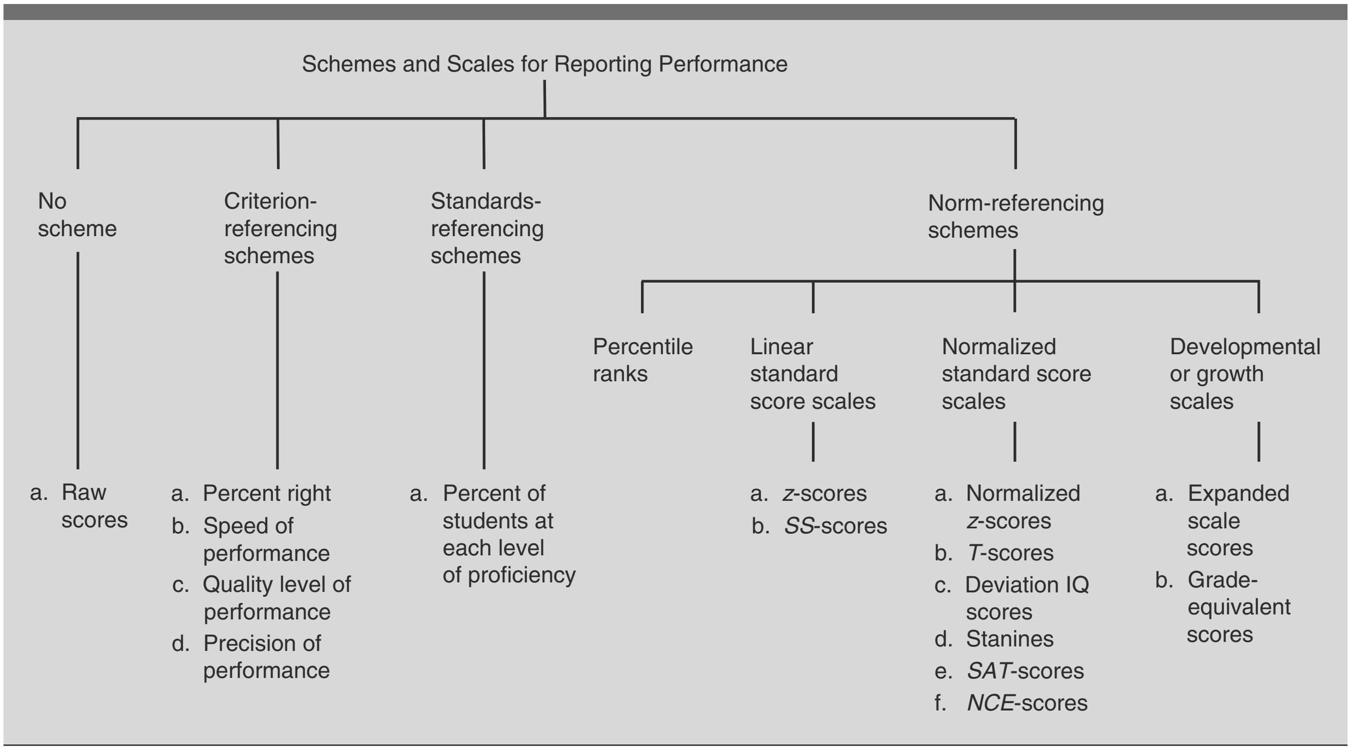

Schemes and Scales for Reporting Performance
No scheme
Criterion-referencing schemes
Standards-referencing schemes
Norm-referencing schemes
Percentile ranks
Linear standard score scales
Normalized standard score scales
Developmental or growth scales
a. Raw scores
a. Percent right
b. Speed of performance
c. Quality level of performance
d. Precision of performance
a. Percent of students at each level of proficiency
a. *z*-scores
b. *SS*-scores
a. Normalized *z*-scores
b. *T*-scores
c. Deviation IQ scores
d. Stanines
e. *SAT*-scores
f. *NCE*-scores
a. Expanded scale scores
b. Grade-equivalent scores

FIGURE 16.3 Example of a percentile norms table: *Differential Aptitude Tests, Level 1, Form C.*

MECHANICAL REASONING											
MALE				FEMALE				COMBINED			
Raw Score	% -ile Rank	Sta- nine	Scaled Score	Raw Score	% -ile Rank	Sta- nine	Scaled Score	Raw Score	% -ile Rank	Sta- nine	Scaled Score
60	99	9	343	60	99	9	343	60	99	9	343
59	99	9	330	59	99	9	330	59	99	9	330
58	99	9	316	58	99	9	316	58	99	9	316
57	99	9	307	57	99	9	307	57	99	9	307
56	99	9	301	56	99	9	301	56	99	9	301
55	99	9	296	55	99	9	296	55	99	9	296
54	99	9	292	54	99	9	292	54	99	9	292
53	98	9	288	53	99	9	288	53	99	9	288
52	97	9	285	52	99	9	285	52	99	9	285
51	95	8	282	51	99	9	282	51	98	9	282
50	94	8	279	50	99	9	279	50	97	9	279
49	92	8	277	49	99	9	277	49	97	9	277
48	90	8	274	48	98	9	274	48	95	8	
47	88	7	272		98	9	272	47			
46	85	7				9	270	46			
45	82	7				9	268				
44	79					9	266				234
				23	15				14	3	232
				22	14				13	3	231
				21	13	2		21	10	2	229
			227	20	12	2		20	8	2	227
19	6	2	226	19	10	2	226	19	7	2	226
18	5	2	224	18	9	2	224	18	5	2	224
17	4	2	222	17	7	2	222	17	4	2	222
16	3	1	220	16	6	2	220	16	3	1	220
15	2	1	218	15	4	1	218	15	3	1	218
14	2	1	216	14	3	1	216	14	2	1	216
13	1	1	214	13	2	1	214	13	2	1	214
12	1	1	212	12	2	1	212	12	1	1	212
11	1	1	209	11	1	1	209	11	1	1	209
10	1	1	207	10	1	1	207	10	1	1	207
9	1	1	204	9	1	1	204	9	1	1	204
8	1	1	201	8	1	1	201	8	1	1	201
7	1	1	198	7	1	1	198	7	1	1	198
6	1	1	194	6	1	1	194	6	1	1	194
5	1	1	190	5	1	1	190	5	1	1	190
4	1	1	185	4	1	1	185	4	1	1	185
3	1	1	178	3	1	1	178	3	1	1	178
2	1	1	169	2	1	1	169	2	1	1	169
1	1	1	155	1	1	1	155	1	1	1	155

Source: From *Differential Aptitude Tests, Fifth Edition, Fall Norms Booklet.* Copyright © 1991, 1989 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

will use the test scores. Be sure to use the norms table that corresponds to the time of year during which the student takes the assessment. In our example, a raw score of 48 in the fall of the year corresponds to a percentile rank of 99. If you looked up a raw score of 48 in the spring norms table, it would have a slightly lower percentile rank. This lower percentile rank reflects that students learn or improve during the year.

As with all scores, you should not interpret percentile ranks too precisely. For example, a student with a percentile rank of 44 and a student with a percentile rank of 46 differ little. Therefore, for many educational decisions you should interpret these scores as essentially equivalent. Some publishers, to reflect that all scores contain measurement error, report percentile bands or uncertainty intervals instead of a single percentile rank. These percentile bands are based on the assessment's standard error of measurement (see Chapter 4).

Percentile ranks have some advantages. Percentile ranks:

- Are easily understood by pupils, parents, teachers, and others.
- Clearly reflect the norm-referenced character of the interpretation.
- Permit a person's performance to be compared to a variety of norm groups.
- Can be used to compare a student's relative standing in each of several achievement or ability areas.

They also have some limitations. Percentile ranks:

- Can be confused with percentage correct scores.
- Can be confused with some other types of two-digit derived scores.
- Do not form an equal-interval scale. Differences between *PR*s in the middle of the scale tend to be overinterpreted. Differences of the same magnitude near the tails of a distribution tend to be underinterpreted.

Because percentile ranks are easy to understand, your school district will most likely report them. Percentile ranks are also easy to calculate. Figure I.8 in Appendix I shows the procedure. The same procedure can be used for results from your classroom or for your entire district.

Remember that percentile ranks are specific to the group being referenced. After your students take a standardized test, the publisher will probably report both the **local percentile ranks** and the **national percentile ranks**. Your student Robert, for example, may have a national percentile rank of 40 and a local percentile rank of 30. Local percentile ranks are lower than national percentile ranks only when the population of students in a local school system scores higher, on the average, than the national standardization sample. Keep the reference group in mind when you interpret percentile ranks.

LINEAR STANDARD SCORES

The second branch of the norm-referencing schemes diagram in Figure 16.2 shows two types of linear standard scores. Both are discussed in this section. A **linear standard score** tells how far a raw score is from the mean of the norm group, the distance being expressed using standard deviation units. The standard deviation is an index that measures the spread of scores in a distribution. The standard deviation is denoted *SD* in this book and is explained in Appendix I.

In general, linear standard scores have the same-shaped distribution as the raw scores from which they are derived (this is not true of percentile ranks and nonlinear standard scores) and can be used to make two distributions more comparable by placing them on the same numerical scale. Linear standard scores are called linear because if you plot each raw score against its corresponding linear standard score in a graph and then connect these points, you will always have a straight line.

z-Scores

The fundamental linear standard score is the ***z*-score**, which tells the number of standard deviation units a raw score is above (or below) the mean of a given distribution. Other linear standard scores are computed from *z*-scores. Equation 16.1 explains.

$$z = \frac{X - M}{SD} \qquad \text{[Eq. 16.1]}$$

where

X represents the raw score

M represents the mean (average) raw score of the group

SD represents the standard deviation of the raw scores for that group

Here is an example of how to apply this equation:

Example

Example of calculating a linear z-score using Equation 16.1

Suppose Ashley's raw score was 38 on Test A. Suppose further that the test mean is 44 and the standard deviation is 4. The corresponding *z*-score is calculated as follows:

$$z = \frac{38 - 44}{4} = \frac{-6}{4} = -1.5$$

The *z*-score tells the number of standard deviations a raw score is above or below the mean. For example, if a student's raw score falls below the mean a distance equal to one and one half times the standard deviation of the group, the student's *z*-score equals −1.5. A *z*-score is negative when the raw score is below the mean, positive when the raw score is above the mean, and equal to zero when the raw score is exactly equal to the mean.

An advantage of using *z*-scores is that they communicate students' norm-referenced achievement expressed as a distance away from the mean. In many groups, the majority of students' scores cluster near the mean, usually within one standard deviation on either side of the mean. A distance of one standard deviation above the mean is $z = +1.0$; a distance of one standard deviation below the mean is $z = -1.0$. Thus, you would interpret a student whose *z*-score is between +1.0 and −1.0 as having typical or average attainment relative to others. Similarly, you would interpret a student with $z = -1.5$ or less as having atypically low attainment because few students have *z*-scores of −1.5 or less. You interpret a student with $z = +1.5$ or greater as having atypically high attainment because relatively few students attain *z*-scores of +1.5 or greater.

Another advantage of using *z*-scores is to put raw scores with different metrics on the same norm-referenced scale. Consider the following example, in which the same students are measured in both pounds and kilograms. Notice what happens when each student's measurements are transformed to *z*-scores.

Example

Example showing how a student's z-scores remain the same even though the measurement scale changes

	Weight in kilograms		*Weight in pounds*	
Student	*X*	*z*	*X*	*z*
A	48	–1.52	105.2	–1.52
B	52	–0.17	114.4	–0.17
C	54	0.51	118.8	0.51
D	56	1.18	123.2	1.18

Even though the pounds mean and standard deviation are different from the kilograms mean and standard deviation, the students' relative positions in the distributions are the same. This is expressed by the *z*-scores (which are identical for pounds and kilograms), not by the pounds and kilograms raw scores.

The *z*-score has several practical disadvantages. It is difficult to explain to students and parents, because understanding it requires an understanding of the mean and standard deviation. Another practical disadvantage is that plus and minus signs are used. Transcription errors, resulting in omitted or interchanged signs, are frequent. Further, you will find it difficult to explain to students (or parents) why assessment performances are reported as negative and/or fractional numbers. For example, a student may say, "I got 15 of the 45 questions right. How could my score be −1.34?" Likewise, the decimal point is subject to frequent transcription error.

These practical problems are easily overcome, however, by transforming the *z*-score to other types of scores. These additional transformations maintain the conceptual norm-referenced advantage of *z*-scores while overcoming their practical limitations.

SS-Scores

The second type of score under the linear standard score branch of Figure 16.2 is the *SS*-score. An ***SS*-score** tells the location of a raw score in a distribution having a mean of 50 and a standard deviation of 10. To remedy some of the disadvantages of *z*-scores, some publishers apply a modification (transformation) to eliminate both the negative scores and the fractional portion of the *z*-scores. Equation 16.2 for an *SS*-score shows how these two things are accomplished:

$$\begin{aligned} SS &= 10z + 50 \\ &= (10 \text{ times the } z\text{-score}) + 50 \quad \text{[Eq. 16.2]} \end{aligned}$$

First, *z*-scores are computed; then, each *z*-score is transformed to an *SS*-score: Each *z* is multiplied by 10, the product rounded to the nearest whole number, and finally 50 is added. Multiplying by 10 and rounding eliminates the *z*-score's decimal. Adding 50 eliminates the *z*-score's negative value. Here is an example of how to use the equation:

Example

***Example showing how a student's z-score is transformed into an* SS-score.**

Suppose Ashley's *z*-score was computed to be –1.5. (See the earlier example.) To convert this to an *SS*-score, multiply it by 10 and add 50 to the result. Thus,

$$\begin{aligned} SS &= 10(-1.5) + 50 \\ &= -15 + 50 = 35 \end{aligned}$$

The result of applying this conversion to the *z*-scores is that the distribution of *SS*-scores will have a mean of 50 and a standard deviation of 10. Once you know this fact, you can interpret anyone's *SS*-score, essentially by doing a mental conversion back to a *z*-score.

Example

Example showing how to interpret a student's* SS*-score by converting it back to a z-score

Ashley's *SS*-score is 35; a score of 35 is 15 points or 1.5 standard deviations below 50, the mean. Thus, Ashley's *z*-score is −1.5.

SS-scores have the advantage of not changing the shape of the original raw score distribution. The distribution of *SS*-scores always has a mean of 50 and a standard deviation of 10. The *SS*-score is interpretable in terms of standard deviation units while avoiding negative numbers and decimal fractions. A disadvantage is that a person needs to understand the concepts of standard deviation and linear transformation to interpret them.

Comparison of Linear Standard Scores

It may help you understand these scores if we display the numerical relationship between them. Because all linear standard score systems reflect essentially the same information, interpreting their meaning is easy once you know the multiplier and the added constant. The next example shows how each type of score is related to the other and to the raw scores:

Example

Example comparing z-scores and* SS*-scores for the same raw score

Raw score in a group with M = 41 and SD = 3	*Linear standard scores corresponding to each raw score*	
	z-score	*SS-score*
32	–3.0	20
35	–2.0	30
38	–1.0	40
41	0.0	50
44	+1.0	60
47	+2.0	70
50	+3.0	80

NORMAL DISTRIBUTIONS

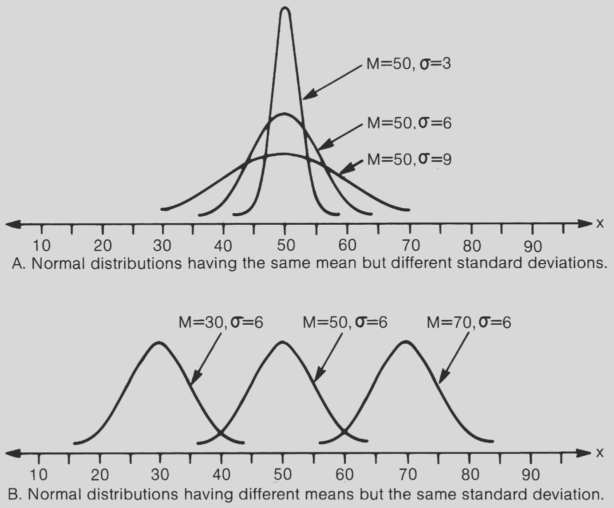
Shortly we will discuss the normalized standard score branch of Figure 16.2. However, first we need to discuss normal distributions of scores.

Definition

Assessment developers have found it advantageous to transform the scores to a common distributional form: a normal distribution. A **normal distribution**, sometimes called a *normal curve*, is a mathematical model invented in 1733 by Abraham de Moivre (Pearson, 1924). It is defined by a particular equation that depends on two specific numbers: the mean and the standard deviation, signifying that many normal distributions exist and each has a different mean and/or standard deviation. Figure 16.4 shows several different normal curves. Each of these was obtained by using the normal curve equation and plotting points on a graph. In Figure 16.4 (A), each normal distribution has the same mean but a different standard deviation. Although each is centered on the same point on the *X*-scale, some appear flatter and more spread out because their standard deviation is larger. Figure 16.4 (B) shows three normal curves, each with the same standard deviation but each with a different mean. The degree of spread is the same for each, but each is centered on a different point on the score scale.

Every normal curve is smooth and continuous; each has a symmetrical, bell-shaped form. In theory, a normal curve never touches the baseline (horizontal axis) but is asymptotic to it, extending out to infinity in either direction from the mean. Graphs of actual raw-score distributions are nonsymmetrical and jagged. For actual raw-score distributions, the lowest possible score is 0 and the highest possible score equals the total number of items on the assessment. An idea of how an actual distribution compares to the mathematically defined normal curve may be obtained from Figure 16.5. Both distributions have the same mean and the same standard deviation. This normal

FIGURE 16.4 Illustrations of different normal distributions.

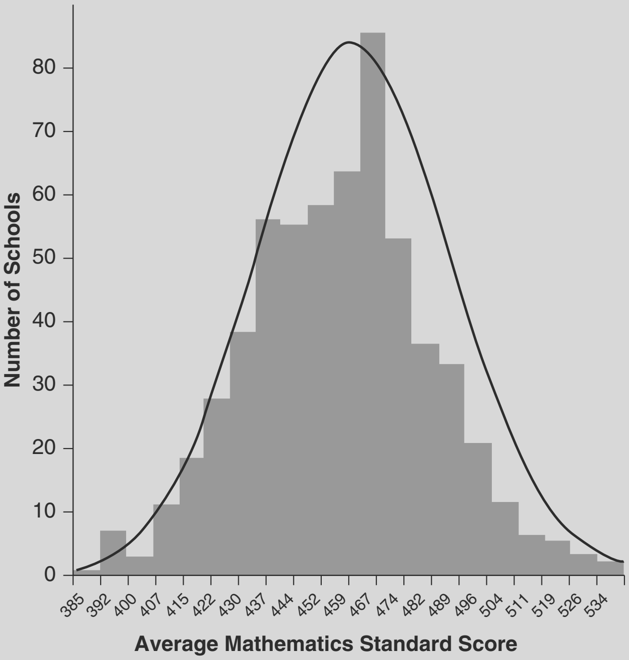


A. Normal distributions having the same mean but different standard deviations.

B. Normal distributions having different means but the same standard deviation.

Source: From *Measuring Pupil Achievement and Aptitude* (2nd ed., p. 87), by C. M. Lindvall and A. J. Nitko, 1975, New York: Harcourt Brace Jovanovich. Reprinted by permission of the authors.

FIGURE 16.5 Example of a mathematically defined normal curve (smooth curve) superimposed on an actual distribution (histogram) of average eighth-grade mathematics standard scores for 575 schools. Both distributions have the same mean and standard deviation.



Source: Histogram drawn from 2004 eighth-grade school averages from the Arizona Department of Education, Accountability Division, Research and Evaluation Section, http://www.ade.state.az.us/profile/publicview/Download.asp. Used with permission.

curve approximates the actual distribution but does not match it exactly.

Natural Law Versus Normal Distributions

Early users of normal curves believed that somehow natural laws dictated that nearly all human characteristics were distributed in a random or chance fashion around a mean or average value. This view of the normal curve's applicability was, perhaps, begun by de Moivre (1756), but it was adamantly held to be true for intellectual and moral qualities by Quetelet (1748) (Dudycha & Dudycha, 1972; Landau & Lazarsfeld, 1968).

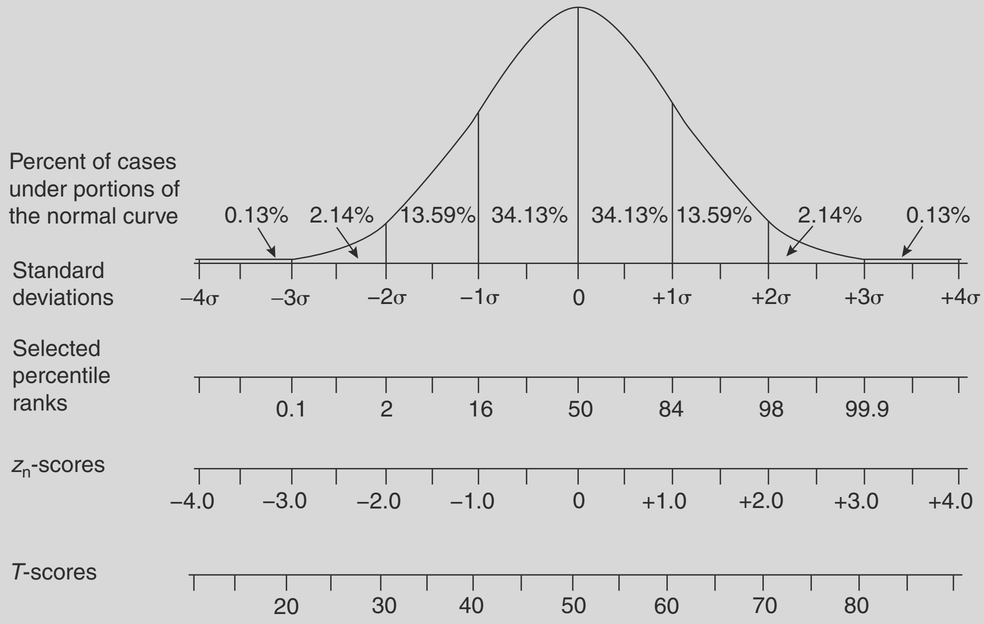
This thought—that somehow the distributions of human characteristics are by nature normal distributions—has carried over to mental measurement. It is frequently held, too, that because assessment scores have a bell-shaped distribution, this indicates that not just the scores but also the human abilities *underlying* the scores are normally distributed. This statement is, of course, not true. The assessment's score distribution depends not only on the underlying abilities of the persons tested but also on the properties of the assessment procedure itself. An assessment developer can, by judicious selection of tasks, make the score distribution have any shape: rectangular, skewed bimodal, symmetrical, and so on (Lord, 1953). (See Appendix I for shapes of distributions.) These nonnormal score distribution shapes could appear in the data, for example, even though the underlying ability of the group is normal in form. Similarly, score distributions could appear to be normal in shape even though the underlying ability of the group is nonnormal in form.

From your own experience, you know that you can control the shape of a test score distribution. For example, if all items on a test are easy, there will be a lot of high scores and few low scores. A very difficult test will have many low scores and few high scores. The point is, the normal distribution is a convenient model, but you should not believe it is a natural representation of educational achievement outcomes.

Percentile Ranks and *z*-Scores in a Normal Distribution

To understand the relationship between percentile ranks and normal curve *z*-scores, look at the graph at the top of Figure 16.6. If we cut up a normal

FIGURE 16.6
Relationships among percentile ranks, *z*-scores, and *T*-scores in a normal distribution.



%ile	Normalized*:		%ile	Normalized*:		%ile	Normalized*:		%ile	Normalized*:	
rank	z_n	*T*	rank	z_n	*T*	rank	z_n	*T*	rank	z_n	*T*
0.05	−2.6	24									
1	−2.3	27	26	−0.6	44	51	0.0	50	76	0.7	57
2	−2.1	29	27	−0.6	44	52	0.1	51	77	0.7	57
3	−1.9	31	28	−0.6	44	53	0.1	51	78	0.8	58
4	−1.8	32	29	−0.6	44	54	0.1	51	79	0.8	58
5	−1.7	33	30	−0.5	45	55	0.1	51	80	0.8	58
6	−1.6	34	31	−0.5	45	56	0.2	52	81	0.9	59
7	−1.5	35	32	−0.5	45	57	0.2	52	82	0.9	59
8	−1.4	36	33	−0.4	46	58	0.2	52	83	1.0	60
9	−1.3	37	34	−0.4	46	59	0.2	52	84	1.0	60
10	−1.3	37	35	−0.4	46	60	0.3	53	85	1.0	60
11	−1.2	38	36	−0.4	46	61	0.3	53	86	1.1	61
12	−1.2	38	37	−0.3	47	62	0.3	53	87	1.1	61
13	−1.1	39	38	−0.3	47	63	0.3	53	88	1.2	62
14	−1.1	39	39	−0.3	47	64	0.4	54	89	1.2	62
15	−1.0	40	40	−0.3	47	65	0.4	54	90	1.3	63
16	−1.0	40	41	−0.2	48	66	0.4	54	91	1.3	63
17	−0.9	41	42	−0.2	48	67	0.4	54	92	1.4	64
18	−0.9	41	43	−0.2	48	68	0.5	55	93	1.5	65
19	−0.9	41	44	−0.2	48	69	0.5	55	94	1.6	66
20	−0.8	42	45	−0.1	49	70	0.5	55	95	1.7	67
21	−0.8	42	46	−0.1	49	71	0.6	56	96	1.8	68
22	−0.8	42	47	−0.1	49	72	0.6	56	97	1.9	69
23	−0.7	43	48	−0.1	49	73	0.6	56	98	2.1	71
24	−0.7	43	49	−0.0	50	74	0.6	56	99	2.3	73
25	−0.7	43	50	−0.0	50	75	0.7	57	99.9	3.1	81

*Values are rounded. To “normalize“ scores, enter table with actual percentile rank and read out z_n or *T*.

distribution into sections one standard deviation wide, each section will have a fixed percentage of cases or **area under the normal curve**. For example, a section that is one standard deviation wide and located just above the mean contains approximately 34% of the area. The comparable section just below the mean contains, by symmetry, 34% as well. Together those two sections contain 68%

of the area. Thus, 68% of the area in a normal distribution will be within one standard deviation of the mean; 95% will be within two standard deviations; and 99.7% of the area will fall within three standard deviations. Therefore, if a distribution is normal, nearly all of the scores will span a range equivalent to six standard deviations.

You can use these facts about the percentage of cases in various segments to determine the correspondence between percentile ranks and *z*-scores in a normal distribution. To emphasize that we are speaking only of a normal distribution, Figure 16.6 denotes the *z*-scores as z_n. This percentile rank correspondence, the same for all normal distributions, permits an easy interpretation of standard scores in normal distributions. For example, look at the graph in Figure 16.6 and the two scales below the graph. The percentage of cases below $z_n = -2.00$ is 2.27% (= 0.13 + 2.14). (Figure 16.6 also shows *T*-scores, which we will explain later in this chapter.) Thus, in a normal distribution the percentile rank corresponding to $z_n = -2.00$ is (rounded) 2. Other z_n-scores' percentile ranks can be computed similarly from Figure 16.6, as shown in the examples below. The chart under the drawing in Figure 16.6 provides more complete information on percentile rank correspondences between z_n-scores and normal curves.

Example

How to determine the percentile rank corresponding to selected z_n-scores in a normal distribution

z_n	*PR (rounded)*	*How calculated*
−3.0	0.1	= 0.13
−2.0	2	= 0.13 + 2.14
−1.0	16	= 0.13 + 2.14 + 13.59
0.0	50	= 0.13 + 2.14 + 13.59 + 34.13
1.0	84	= 50 + 34.13
2.0	98	= 50 + 34.13 + 13.59
3.0	99.9	= 50 + 34.13 + 13.59 + 2.14

NORMALIZED STANDARD SCORES

Now that you have a little background on the meaning of a normal curve, let's return to the third branch of Figure 16.2: normalized standard scores. The figure shows five types of normalized standard scores. We will discuss all of them in this section.

Test publishers may transform raw scores to a new set of scores that is distributed normally (or nearly so). Such transformation changes the shape of the original distribution, squeezing and stretching the scale to make it conform to a normal distribution. Once this is accomplished, various types of standard scores can be derived, and each can have an appropriate normal curve interpretation. The general name for these derived scores is **normalized standard scores**. These are also termed *area transformations,* as opposed to linear transformations, which we presented earlier in this chapter. This section reviews five of the common varieties reported in test manuals and shown in Figure 16.2.

Normalized *z*-Scores

When the *z*-scores have percentile ranks corresponding to what we would expect in a normal distribution, they are called normalized *z*-scores, **or z_n-scores,** and the following symbol is used:

z_n = the *z*-score corresponding to a given percentile rank in a normal distribution

If a distribution of raw scores is not normal in form, the percentile ranks of its *z*-scores will not correspond to what would be expected in a norm distribution. You may be surprised to learn, however, that one can create a set of "normalized" *z*-scores for any nonnormal distribution. After making this transformation, the new set of scores is more nearly like a normal distribution. **Normalizing a set of scores** is done in the following way: (a) determine the percentile rank of each raw score in the norm group, (b) look up each percentile rank in a normal curve table (e.g., the chart in Figure 16.6), and (c) read out the z_n-value that corresponds to each. The resulting z_n-values are "normalized." That is, they are the *z*-scores that *would have been attained if the distribution had been normal in form.*

To show you how the process works, and to illustrate the difference between *z* and z_n, consider the scores in the next example. The scores and the percentile ranks came from our example of the class of 25 students that showed how percentile ranks were calculated (Figure I.8 in Appendix I).

Example

Illustration of normalized z-scores and (actual) linear z-scores corresponding to the distribution of 25 test scores shown in the previous example in Figure I.8

Raw score	*Percentile rank*	*Normalized[a] standard scores (zn)*	*Linear[b] standard scores (z)*
36	98	2.05	2.43
33	96	1.75	1.64
32	94	1.55	1.38
31	90	1.28	1.12
30	88	1.18	0.86
29	84	0.99	0.59
28	72	0.58	0.33
27	54	0.10	0.07
26	32	–0.47	–0.20
25	16	–0.99	–0.46
24	10	–1.28	–0.72
22	8	–1.41	–1.25
21	6	–1.55	–1.51
15	4	–1.75	–3.09
14	2	–2.05	–3.36

Notes: [a]z_n-values are obtained by looking up the percentile ranks in Figure 16.6 and reading out the corresponding z_n-values.

[b]z-values are obtained by using the actual distribution of scores in Table I.8 (Appendix I) and by applying the equation:

$$z = \frac{X - M}{SD}$$

where $M = 26.75$ and $SD = 3.80$.

Next, you look up each percentile rank in Figure 16.6, and read out the corresponding z_n. The results appear in the example. For the sake of comparison, the actual, linear z-scores are computed via Equation 16.1, using $M = 26.75$ and $SD = 3.8$. The difference between the normalized and linear z-scores represents the "stretching and squeezing" necessary to make the original distribution correspond more nearly to a normal distribution.

Normalized *T*-Scores (McCall's *T*)

The second type of score in the normalized standard score branch of Figure 16.2 is a ***T*-score**. A normalized *T*-score tells the location of a raw score in a normal distribution having a mean of 50 and a standard deviation of 10. The normalized *T*-score is the counterpart to the linear *SS*-score. Thus,

$$T = 10z_n + 50 \qquad \text{[Eq. 16.3]}$$

The difference between Equation 16.3 and Equation 16.2 ($SS = 10z + 50$) is that z_n is a normalized standard score instead of a linear standard score.

Normalized *T*-scores have the same advantages over normalized *z*-scores as *SS*-scores have over linear *z*-scores, with the additional advantage that *T*-scores have the percentile rank interpretations of a normal curve. Here is an example:

Example

Examples of how to interpret T-scores using a normal curve like the one shown in Figure 16.6

1. Joey's *T*-score is 40. This means he is one standard deviation below the mean of the norm group, and his percentile rank is approximately 16.
2. Keisha's percentile rank is 84. This means her *T*-score is 60, and she is a distance of one standard deviation above the norm-group mean.

Figure 16.6 shows the correspondence between percentile ranks, *T*-scores, and z_n scores in a normal distribution. That figure can help you convert percentile ranks directly to *T*-scores without using Equation 16.3.

Deviation IQ Scores

The third type of normalized standard score shown in Figure 16.2 is the deviation IQ score used with certain assessments of mental ability. A **deviation IQ score,** or **DIQ-score,** tells the location of a raw score in a normal distribution having a mean of 100 and a standard deviation of 15 or 16. The norm group is usually made up of all those students with the same chronological age, regardless of grade placement. For example, if the test developer sets the standard deviation at 16, *DIQ*s are given by

$$DIQ = 16z_n + 100 \qquad \text{[Eq. 16.4]}$$

These *DIQ*s are interpreted in a way similar to *T*-scores, but with reference to the normal distribution having a mean of 100 and a standard deviation of 16. Here is an example:

Example

***The meaning of* DIQ-scores**

1. Meghan has *DIQ* = 116. This means she has scored one standard deviation above the mean of her age group and the percentile rank of her score is 84.
2. Sherry has *DIQ* = 100. This means she has scored at the mean of her age group and the percentile rank of her score is 50.

Usually, assessment manuals provide tables that permit you to convert raw scores directly to *DIQ*s.

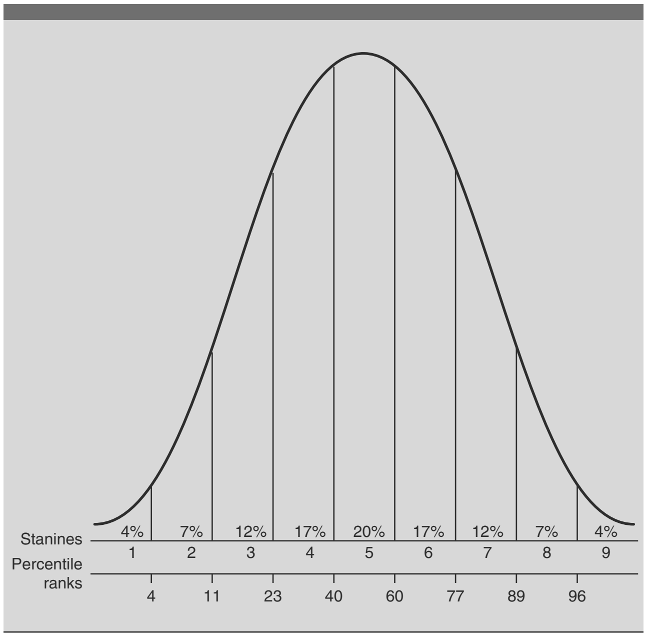
Stanines

The fourth normalized standard score shown in Figure 16.2 is the stanine. A **stanine score** tells the location of a raw score in a specific segment of a normal distribution. Publishers frequently recommend using **national stanines** for norm-referenced interpretation of achievement and aptitude assessments.

Figure 16.7 illustrates the meaning of a stanine score. A normal distribution is divided into nine segments, numbered from a low of 1 through a high of 9. Scores falling within the boundaries of these segments are assigned one of these nine numbers (hence, the term *stanine* from "standard nine"). Each segment is one half a standard deviation wide, except for stanines 1 and 9. The percentage of the cases in a normal curve falling within each segment is shown in Figure 16.7, along with the range of percentile ranks associated with each.

All persons with scores falling within an interval are assigned the stanine of that interval. For example, all persons with scores having percentile ranks from 11 through 22 are assigned a stanine of 3; all from 23 through 29 a stanine of 4; and so on.

FIGURE 16.7 Illustration of a normal distribution showing stanines, percentile ranks, and percentage of cases having each stanine.



Twelve percent of the persons in the norm group would be assigned a stanine of 3 and 17% a stanine of 4. When raw scores from normal distributions are converted to stanines, the stanines have a mean of 5 and a standard deviation equal to 2. Here is an example of how stanines are interpreted. As you read these examples, refer to Figure 16.7.

Example

How to interpret stanine scores

1. Sophia received a stanine of 5 on the mathematics subtest of a standardized test. This means her raw score on the test was in the middle 20% of the norm group.
2. Jesse received a stanine of 9 in the reading subtest of a standardized test. This means his raw score on the test was in the top 4% of the norm group.
3. Blake's stanine on the spelling subtest of the standardized test was 3. This means that his raw score was in the lower 20% of the norm group. Specifically, his percentile rank was between 11 and 22.

Among the advantages claimed for stanines: They are always single-digit numbers, have approximately equal units all along the score scale, and do not imply an exactness greater than that warranted by the assessment.

Not all assessment experts agree with using stanines for norm-referenced interpretations. Some hold that stanines present more difficult interpretative problems than percentile ranks, especially for reliable assessments, because stanines reflect coarse groupings of scores.

As with percentile ranks, stanines are specific to the reference group on which they are calculated. Some test publishers report both local and national stanines. For a specific student, these two stanines may be different, depending on how the student ranks in each reference group.

The example in Appendix I (Figure I.11) shows how you can transform any set of scores into stanines. The example uses the distribution of 25 scores we used earlier in Figure I.8.

SAT-Scores

The fifth score in the normalized standard score branch of Figure 16.2 is the ***SAT*-score**. The *SAT Reasoning Test* (SAT) results are reported using this type of score. The *SAT*-score is a normalized standard score from a distribution that has a mean of 500 and a standard deviation of 100. The *SAT*-score scale is based on a reference group of 1,052,000 students who graduated from high school in 1990 and who took the *SAT* in either their junior or senior year. The scores were recently recentered (Dorans, 2002), using transformations that are beyond the scope of this book. The purpose of the recentering, however, was to bring scores back into line so that they would communicate the original meaning of the scores developed from the 1990 reference group. The scores of this reference group were normalized, the mean set to 500, and the standard deviation set to 100. This is shown in Equation 16.5:

$$SAT\text{-score} = 100z_n + 500 \qquad \text{[Eq. 16.5]}$$

Tests are statistically equated to the recentered scores that were based originally on the 1990 reference group. This ensures that the scores have the same meaning from year to year. Percentile ranks corresponding to each current year's scores are provided to test users to facilitate interpretation for the current year.

Normal Curve Equivalents

The sixth type of normalized standard score in Figure 16.2 is the normal curve equivalent. The **normal curve equivalent (NCE)** is a normalized standard score with a mean of 50 and a standard deviation of 21.06. It was developed primarily for use with federal program evaluation efforts (Tallmadge & Wood, 1976). Its primary value is evaluating gains from various educational programs that use different publishers' tests. *NCE*-values are found by the formula shown in Equation 16.6. Their highest possible value is 99 and their lowest possible value is 1.

$$NCE = 21.06z_n + 50 \qquad \text{[Eq. 16.6]}$$

As stated previously, *NCE*-scores have a mean of 50 and a standard deviation of 21.06. By comparison, *T*-scores have a mean of 50 and a standard deviation of 10. Why choose a standard deviation of 21.06? This choice of standard deviation was made so the *NCE*-scores would span the range 1 to 99.

The following example shows the relationship between selected percentile ranks, *NCE*-scores, and stanines:

Example

Correspondences between selected percentile ranks, NCE-scores, and stanines

Percentile rank	*NCE*	*Stanine*
1	1	1
5	15	2
10	23	2
20	32	3
25	36	4
30	39	4
35	42	4
40	45	5
45	47	5
50	50	5
55	53	5
60	55	6
65	58	6
70	61	6
75	64	6
80	68	7
85	72	7
90	77	8
95	85	8
99	99	9

As you can see in the table, percentile ranks of 1, 50, and 99 are identical in value to *NCE*-scores. At other points, however, percentile ranks and *NCE*-scores differ: *NCE*-scores are less spread out than percentile ranks in the middle of the distribution and more spread out than percentile ranks at the lower and upper extremes. Notice the *NCE*-scores look very similar to percentile ranks. This is why they are often confused with percentile ranks. Although some publishers present *NCE* norms tables in their standardized test manuals, we do not recommend *NCE*-scores for reporting individual student results because they are too easily confused with percentile ranks.

You may notice the relationship of the *NCE*-score to stanines. If you move the *NCE* decimal point to the left one digit and round to the nearest whole number, you will roughly have the stanine. For example, an $NCE = 72$ has a stanine equivalent of 7; $NCE = 58$ has a stanine equivalent of 6; and so on. This rough correspondence stems from the fact that both *NCE*-scores and stanines are based on a normal distribution, and *NCE*-scores and percentile ranks have the same range.

DEVELOPMENTAL AND EDUCATIONAL GROWTH SCALES

We turn now to the fourth branch of the norm-referencing schemes in Figure 16.2. The normalized standard score scales discussed so far are specific to a particular grade level or age group. If a score scale is specific to a particular grade, you cannot use it to measure growth as a student moves from one grade to the next. For example, suppose Billy tested at the 84th percentile in Grades 5, 6, and 7. Although Billy would be growing in skills and knowledge, his percentile rank (84) has stayed the same. The number, 84, by reflecting only location in each grade's norm group, does not communicate Billy's growth. Similarly, suppose Ashley's *T*-score determined separately for each grade's norm group remained nearly the same from year to year, say about 60. Ashley in fact exhibited educational growth each year as she moved through the grades. The *T*-score, because it remains constant, does not communicate growth.

You would find it useful, however, if your students' educational growth were reported on one scale of numbers that spanned the school years. Survey achievement batteries, for example, usually span several grades—say 2nd through 8th, or 9th through 12th. If the score scale of such batteries linked the assessments from several grade levels to a single developmental score scale, you could measure your students' growth over those years. We now turn to a discussion of the two scales shown in the developmental or growth scales branch of Figure 16.2: the extended normalized standard score scale and the grade-equivalent score scale.

EXTENDED NORMALIZED STANDARD SCORE SCALES

Basic Idea of the Extended Normalized Score Scales

An **extended normalized standard score** tells the location of a raw score on a scale of numbers that is anchored to a lower grade reference group. Educators find that a "ruler" or achievement continuum on which a student's progress can be measured over a wide range of grades is very useful. On this continuum, low scores represent the lowest levels of educational development and high scores the highest level of educational development. Publishers refer to this type of scale with a variety of names, for example: *obtained scale score, scale score, extended standard score, developmental standard score,* or *growth-scale values*.

Development of Extended Score Scales

Although each publisher prepares expanded scales somewhat differently, and the numbers obtained are not comparable from publisher to publisher, extended scaled scores share the same goals and the same general method of development: (a) a base or anchor group is chosen and normalized *z*-scores are developed that extend beyond the range of scores for this anchor group; (b) a series of assessments are administered with common items given to adjoining groups (e.g., second and third graders take a common set of items, then third and fourth graders, and so on); (c) distributions of scores are tabulated and normalized for each grade; and (d) through these overlapping items, all of the groups are placed on the extended *z*-score scale of the anchor group. This extended *z*-scale becomes the ruler or growth scale spanning the several grades.

The extended *z*-scale is then transformed again to a scale that removes the unpleasant properties (such as negative numbers and decimals) of the *z*-scores. The new scale may range from 00 to 99, from 000 to 999, or any other set of positive integers, depending on the publisher; there are no standards for what this range should be.

Item Response Theory Method

Recent technical advances have given publishers two choices of how extended standard scores can be calculated. One method uses the traditional raw score for students (i.e., number right score) as a beginning step for calculating. This is the method we described above. A second method uses **item response theory (IRT)** in which a mathematical equation is fit to the publisher's sample of students' item responses. The results are then used to derive a score scale. According to this method, students' scores depend on the pattern of their right or wrong answers. **IRT pattern scoring** considers whether students answered an easy or difficult item correctly, and how sharply that item distinguishes students of different achievement levels. This means, for example, that two students who answer correctly the same number of items may get different scaled scores if the pattern of correctly answered items is substantially different for the two students.

The advantage is that the resultant extended standard scores have lower measurement error and greater reliability than traditional number-right scores. The disadvantages are (a) the direct link between the number correct and the extended scale score is broken (when certain equations are used), and (b) the method does not work well for every type of test and every population of students.

A full explanation of item response theory is beyond the scope of this book. To learn more about item response theory you may consult http://edres.org/irt for links to different tutorials. On that site you will also find an introductory book, *The Basics of Item Response Theory* (Baker, 2001), that you may read online.

Recommendations

Although program evaluators and school researchers generally prefer to use extended standard scores, their meaning is not immediately apparent to teachers, parents, and students. To understand what they mean you have to compare a student's score with the average score of students in that grade. Some educators consider this an advantage because it lessens the chance of overinterpreting scores. On the other hand, if no one knows what they mean, they will not be used, and therefore the scores will be underutilized.

Extended standard scores tend to show that on the average students exhibit less achievement growth in the upper elementary grades than in the lower grades. Note that extended standard scores show different standard deviations for school subjects and progressively increasing standard deviations as grade levels increase. Thus you cannot compare a student's extended assessment score from one subject area to another. In this respect, they share a common property with grade-equivalent scores, discussed next.

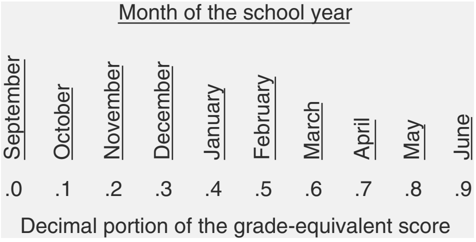
GRADE-EQUIVALENT SCORES

Basic Idea of Grade-Equivalent Scores

A **grade-equivalent score (*GE*)** tells the grade placement at which a raw score is average. *GE*s are the educational development scores most often used with achievement tests at the elementary school level. A grade-equivalent score is reported as a decimal fraction, such as 3.4 or 7.9. The whole number part of the score refers to a grade level, and the decimal part refers to a month of the school year within that grade level. For example, you read a grade-equivalent score of 3.4 as "third grade, fourth month"; similarly, you read 7.9 as "seventh grade, ninth month." Suppose 6.3 is the grade-equivalent score corresponding to the raw score 31. This means that the average in the norm group during the third month of sixth grade was 31. The example below shows how the grade-equivalent scale is laid out:

Example

The grade-equivalent score scale layout



When using these scores, you assume that the time between June and September (i.e., the summer months) represents an increment of one tenth (or 1 month) on the grade-equivalent scale (see above). By defining the grade-equivalent scale this way, the average of the scores shows 10 months' growth every year. However, you should not expect every student to show 1 month's growth each summer.

Overall Usefulness of Grade-Equivalent Scores

Grade-equivalent scores are useful for reporting a pupil's educational development. If a standardized test is administered periodically throughout a student's school years, the resulting grade-equivalent scores can help monitor the student's educational progress using a grade-based educational development scale. To a lesser extent, grade equivalents can be used to evaluate a student's grade placement. The problem with grade placement interpretations of *GE* scores is that they depend on how well the test content matches what was taught to the students up to the point at which the test was administered—the poorer this match, the less valid are grade placement interpretations.

Grade-equivalent scores (and extended standard score) cannot be used to compare a student's strengths and weaknesses across different subject matters. Nor can they be used to determine a student's rank among his or her peers. An explanation of these limitations follows.

Development of Grade-Equivalent Scores

Understanding how the test publisher obtains grade equivalents will help you avoid misinterpretation. There is no need, of course, for you to compute them because test manuals provide the needed conversion tables.

The development process is illustrated with a reading test, but the same process applies to all subject areas. Suppose a publisher wishes to assess reading from Grades 1 through 8 and develop grade equivalents. The publisher creates a series of overlapping tests that spans the desired grades: one test for first and second grades, one for second and third, and so on. Each test is appropriate for specific grade levels. The publisher administers the appropriate tests to a large national sample at each grade level. Usually, the publisher does this once or twice during the year (fall and/or spring) because it is impossible to administer them continuously throughout the year. The dates on which tests are administered are called **empirical norming dates**. These overlapping tests are then linked using an expanded score scale. This allows the raw scores from the different tests to be placed on a grade-based reading ability scale. The process is called *vertical linking* or *vertical equating* because the links go up the grades.

On this common scale, large differences in reading ability exist in the norm group at each grade level. Therefore, at each grade level there is a spread of reading scores. These distributions of reading scores are shown in Figure 16.8. In this illustration, the publisher administered the assessments only once during the year—in February (Month = 0.5)—so the figure graphs the distributions directly above 1.5, 2.5, 3.5, and so forth.

The *GE* is the median score (the mean score is used sometimes instead) in each grade's norm group. Actual grade equivalents can be obtained only for those points in time when the publisher administered the tests. Grade equivalents for other points are obtained by **interpolation or extrapolation**. Interpolation means finding an unknown number between two known numbers (e.g., between the first and second graders' median performance at the norming month). Extrapolation means estimating an unknown number that lies outside the range of available data. Sometimes extrapolation leads to silly interpretations. We once had a parent who wondered why her 7th-grade son received a *GE*-score of 12.0 on a science test in an achievement battery. He had told his parents that he didn't need to pay attention in science class any more, which was definitely not a good interpretation! "How would you expect a 12th grader to score on a 7th-grade science test?" we asked. "He just got all those 7th-grade science questions right."

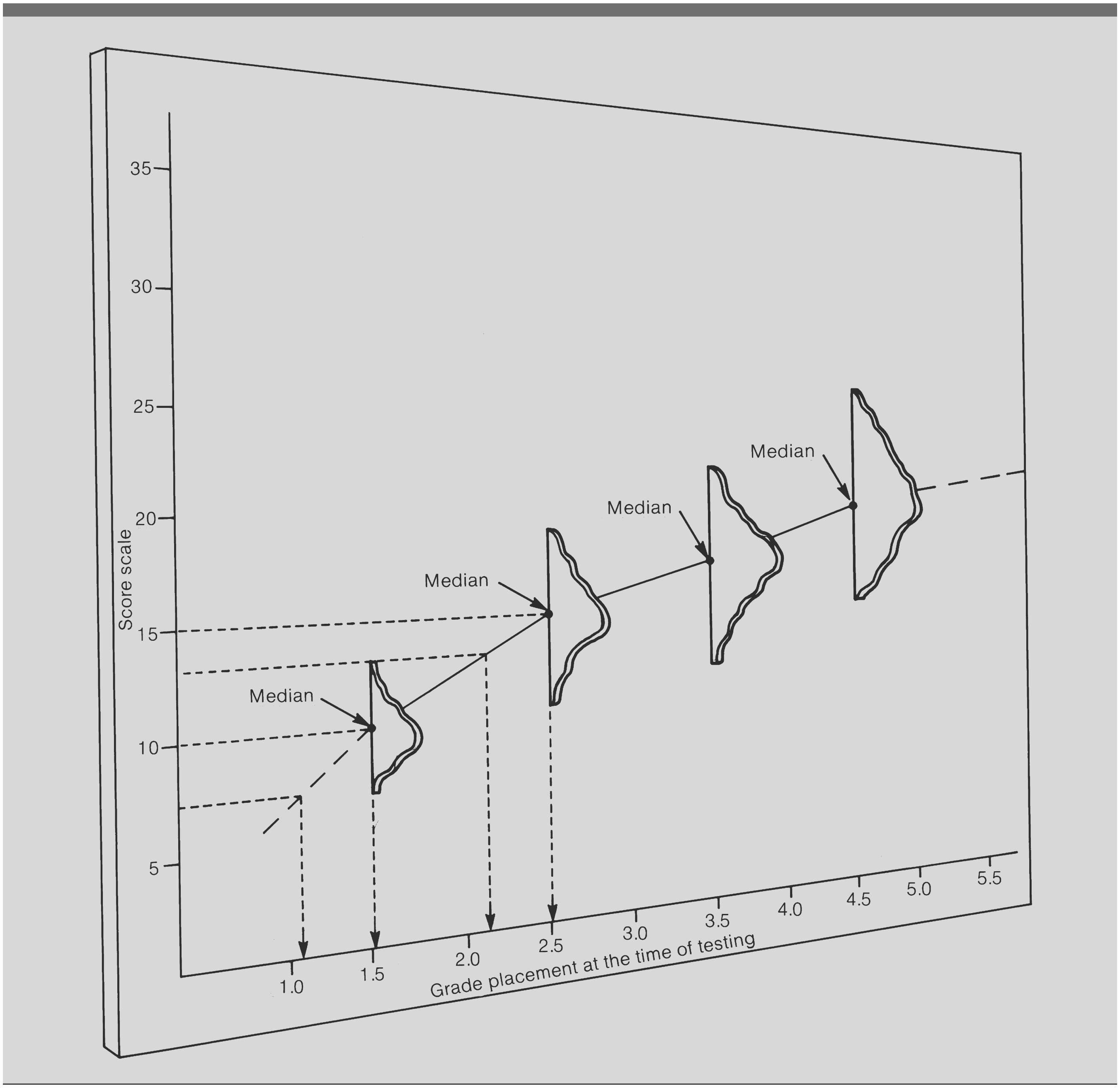
An example of an actual conversion table from a published test is shown in Figure 16.9. You will use this type of conversion table when you consult a publisher's norms booklet to convert your students' raw scores to grade equivalents.

What to Keep in Mind When Interpreting Grade Equivalents

Spring-to-Fall Drops: Summer Losses One special concern in the process of interpreting grade equivalents is the phenomenon of summer achievement losses. In some subject areas—arithmetic, for example—students' performance loses some of its edge over the summer months. A performance drop over the summer months has several meanings: (a) the assumption of an over-the-summer growth of one month is not true in every subject area, (b) educational growth is not regular and uniform for many children, and (c) using fall-to-spring gains in grade-equivalent scores to evaluate an instructional program may lead to wrong conclusions. The third point is less problematic when the test publisher has separate fall and spring norms and when a school system tests on dates very close to the dates on which the publisher's norms were established.

Grade Equivalents and Curriculum Correspondence It would be a misconception to say that students ought to have the same placement as their grade-equivalent scores. To understand why, recall that grade equivalents are based on the median. By definition, half the students in the norm groups at a particular grade placement will have scores above the median. Thus, half the students in the norm group have grade-equivalent scores higher than their actual grade placement. Second, recall that a publisher uses a series of tests, rather than a single test, to establish grade equivalents. You can't interpret a third grader's grade-equivalent score of, say, 5.7 on a mathematics test covering third-grade content to mean that this student ought to be placed in fifth-grade mathematics. The test shows that the student did very well on third-grade content, but the student was not assessed on fifth-grade mathematics. Many factors, of course,

FIGURE 16.8 Hypothetical example of data used to obtain grade-equivalent scores.



besides a single assessment result determine whether the student should receive an accelerated placement. Some test publishers, however, may develop their test batteries so that third-grade students are administered fifth-grade content for purposes of developing a grade-equivalent scale. In such cases, it may be appropriate to say cautiously that the third graders with a grade equivalent of 5.7 do know some fifth-grade content (Hoover et al., 1993b).

The meaning of grade-equivalent scores as describing a student's learning status in a subject depends very much on the subject matter. In reading, for example, students' educational growth may be less tied to the curriculum sequence than it is mathematics. In such cases, third-grade students with grade equivalents of 5.7 may well be reading much like a fifth grader; and a fifth grader with a grade equivalent of 3.7 may well be reading like a third grader.

FIGURE 16.9 **Part of grade-equivalent table for the *Stanford Achievement Tests* (10th Edition), *Forms S/T, Fall Norms.* Notice that you enter the table with the (expanded) scaled scores and read the grade-equivalent scores in the margins.**

						SCALED SCORES					
Grade Equivalent	WORD STUDY SKILLS	WORD RDG/ READING VOCABULARY	SENTENCE READING	READING COMPRE-HENSION	TOTAL READING	MATHEMATICS PROBLEM SOLVING	MATHEMATICS PROCEDURES	TOTAL MATHEMATICS	LANGUAGE MECHANICS A/B	LANGUAGE EXPRESSION A/B	Grade Equivalent
7.9	684	692	680	676	681	674	-	678	672	672	7.9
7.8	683	691	679	675	680	673	685	677	671	671	7.8
7.7	682	690	678	-	679	672	684	676	670	670	7.7
7.6	-	689	-	674	-	671	683	675	669	669	7.6
7.5	681	688	677	-	678	669 – 670	-	674	668	-	7.5
7.4	680	687	676	673	677	668	682	673	667	668	7.4
7.3	-	685 – 686	-	-	-	667	681	672	-	667	7.3
7.2	679	684	675	672	676	666	680	671	666	666	7.2
7.1	678	683	674	-	675	665	679	670	665	665	7.1
7.0	-	682	-	671	-	663 – 664	-	669	664	-	7.0
6.9	677	681	673	-	674	662	678	668	663	664	6.9
6.8	676	680	672	670	673	. 661	677	667	662	663	6.8
6.7	674 – 675	678 – 679	671	669	671 – 672	659 – 660	676	666	661	662	6.7
6.6	673	677	669 – 670	667 – 668	670	658	674 – 675	664 – 665	660	661	6.6
6.5	672	676	668	666	669	657	673	663	659	659 – 660	6.5
6.4	670 – 671	674 – 675	666 – 667	665	667 – 668	655 – 656	672	662	658	658	6.4
6.3	669	673	665	663 – 664	666	654	670 – 671	660 – 661	657	657	6.3
6.2	667 – 668	671 – 672	663 – 664	662	664 – 665	652 – 653	669	659	656	656	6.2
6.1	666	670	662	661	663	651	668	658	655	655	6.1
6.0	665	669	661	659 – 660	662	650	666 – 667	656 – 657	654	653 – 654	6.0
5.9	663 – 664	667 – 668	659 – 660	658	660 – 661	648 – 649	665	655	653	652	5.9
5.8	661 – 662	666	658	657	659	646 – 647	663 – 664	653 – 654	652	651	5.8
5.7	660	664 – 665	656 – 657	655 – 656	657 – 658	644 – 645	660 – 662	651 – 652	650 – 651	649 – 650	5.7
5.6	659	662 – 663	655	653 – 654	656	642 – 643	658 – 659	649 – 650	648 – 649	648	5.6
5.5	657 – 658	661	653 – 654	652	654 – 655	640 – 641	655 – 657	646 – 648	647	647	5.5
5.4	656	659 – 660	652	650 – 651	653	638 – 639	652 – 654	644 – 645	645 – 646	645 – 646	5.4
5.3	655	658	650 – 651	648 – 649	651 – 652	635 – 637	650 – 651	641 – 643	643 – 644	644	5.3
5.2	654	656 – 657	649	646 – 647	650	633 – 634	647 – 649	639 – 640	641 – 642	642 – 643	5.2
5.1	653	654 – 655	647 – 648	645	648 – 649	631 – 632	644 – 646	637 – 638	640	641	5.1
5.0	651 – 652	653	646	643 – 644	647	629 – 630	642 – 643	634 – 636	638 – 639	640	5.0

FIGURE 16.9 (*continued*)

						SCALED SCORES					
Grade Equivalent	WORD STUDY SKILLS	WORD RDG/ READING VOCABULARY	SENTENCE READING	READING COMPRE-HENSION	TOTAL READING	MATHEMATICS PROBLEM SOLVING	MATHEMATICS PROCEDURES	TOTAL MATHEMATICS	LANGUAGE MECHANICS A/B	LANGUAGE EXPRESSION A/B	Grade Equivalent
4.9	650	651 – 652	644 – 645	641 – 642	645 – 646	627 – 628	639 – 641	632 – 633	636 – 637	638 – 639	4.9
4.8	649	650	643	640	644	625 – 626	637 – 638	630 – 631	635	637	4.8
4.7	648	648 – 649	642	-	643	624	636	629	634	636	4.7
4.6	-	646 – 647	641	639	642	623	634 – 635	628	-	635	4.6
4.5	647	644 – 645	640	638	641	-	633	627	-	634	4.5
4.4	646	642 – 643	639	-	640	622	632	626	633	-	4.4
4.3	-	641	638	637	639	621	630 – 631	625	-	633	4.3
4.2	645	639 – 640	637	-	638	620	629	624	-	632	4.2
4.1	644	637 – 638	636	636	637	619	628	623	632	-	4.1
4.0	-	635 – 636	635	635	636	-	626 – 627	622	-	631	4.0

Source: From the 2007 *Fall Supplemental Multilevel Norms Book: Stanford Achievement Test: Tenth Edition*. Copyright © 2007 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

Grade Equivalents and Mastery Sometimes teachers, parents, and school administrators misinterpret grade equivalents as meaning mastery of a particular portion of a curricular area. For example, a parent may erroneously think that a student's grade equivalent of 3.5 in mathematics means the student has mastered 5/10ths of the local school's third-grade mathematics curriculum. The most that can be said, however, is that this student's test score equals the average score of the norm group when it was in the 5th month of third grade. This is unlikely to mean mastery of third-grade mathematics because the test does not systematically sample the entire domain of third-grade mathematics in the student's local curriculum.

Grade Equivalents and What Was Covered in the Class The more closely the test items match the material you emphasized in the classroom before the test was administered, the more likely your students are to score well above grade level on these nationally standardized tests. You may teach the content of some test items after the testing date. As a result, your students may perform poorly when tested but will learn the material before the end of the school year. Answering three or four items wrong will significantly lower a student's grade-equivalent score. If your teaching sequence and the testing sequence are not aligned, inferring mastery is problematic. This points out the norm-referenced character of grade-equivalent scores and illustrates that criterion-referenced interpretations are difficult to make from them.

Grade Equivalents From Different Tests Cannot Be Interchanged Grade-equivalent (and other norm-referenced) scores depend on the particular items placed on the test and the particular norm group used. You would be misinterpreting grade equivalents, for example, if you said, "A grade equivalent of 3.7 on the *ABC Reading Assessment* means the same thing as a grade equivalent of 3.7 on the *DEF Reading Assessment*." The results from two different publishers' assessments are simply not comparable except under special conditions (Peterson et al., 1989).

Grade Equivalents for Different Subjects Cannot Be Compared Another misinterpretation is to compare a student's mathematics grade equivalent with the student's reading grade equivalent. This is invalid. Consider the following hypothetical assessment results for three third-grade students.

Example

		Survey subtest	
		Reading	***Mathematics***
Ian	*GE*	4.9	4.9
	PR	78	90
Santos	*GE*	4.9	4.3
	PR	78	78
Priya	*GE*	4.9	4.6
	PR	78	84

Notice that Ian has two identical grade equivalents, but their corresponding percentile ranks are *different*. Santos has two different grade equivalents but has *identical* percentile ranks. Finally, Priya has one grade equivalent higher than another, yet her higher grade equivalent has a lower percentile rank than her *lower* grade equivalent.

The reason for the phenomena is that scores for one subject area are more diverse than those of another, resulting in different patterns of interpolation when grade equivalents are prepared. Expanded standard scores cannot be used to compare a student's performance in different areas, either.

What should you use to describe a student's relative strengths and weaknesses in different subject areas? *Use percentile ranks to compare a student's scores from different subjects* if all students in the norm group took the same tests in all subjects. Thus, in the preceding illustration, Ian is somewhat better in mathematics and in reading, Santos is about the same in both subjects, and Priya is slightly better in mathematics than in reading. Because these are norm-referenced interpretations, "better" implies "compared with other persons."

"Normal" Growth Sometimes teachers and school administrators use grade equivalents to answer questions of what educational growth they should expect of a student. This is not a good practice and the results of doing it are unsatisfactory. One view of **normal growth** is this: "A student ought to exhibit a growth of 1.0 grade-equivalent unit from one grade to the next." Under this view, a student taking the test in second grade and scoring 1.3, for example, would need to score 2.3 in third grade, 4.3 in fifth, and so on to show "normal" or expected growth.

This *grade-equivalent view of normal growth* cannot be supported at all percentile ranks. Figure 16.10 shows examples of what will happen to three

FIGURE 16.10 **Examples of changes in the percentile ranks for three hypothetical students as each "gains" one year in grade-equivalent units from second through eighth grade.**

Grade placement at the time of testing	*Stanford Achievement Tests,* Total Mathematics						*Iowa Tests of Basic Skills,* Total Mathematics					
	Student A: "Below grade level"		Student B: "On grade level"		Student C: "Above grade level"		Student A: "Below grade level"		Student B: "On grade level"		Student C: "Above grade level"	
	GE	PR	GE	PR	GE	PR	GE	PR	GE	PR	GE	PR
2.3	1.3	25	2.3	61	3.3	86	1.3	12	2.3	60	3.3	93
3.3	2.3	25	3.3	59	4.3	77	2.3	18	3.3	55	4.3	86
4.3	3.3	32	4.3	54	5.3	68	3.3	24	4.3	55	5.3	80
5.3	4.3	36	5.3	52	6.3	69	4.3	29	5.3	54	6.3	74
6.3	5.3	37	6.3	56	7.3	68	5.3	34	6.3	53	7.3	69
7.3	6.3	41	7.3	54	8.3	62	6.3	36	7.3	52	8.3	65
8.3	7.3	43	8.3	52	9.3	59	7.3	39	8.3	51	9.3	63

Source: Data reproduced from the *2007 Fall Supplemental Multilevel Norms Book: Stanford Achievement Test: Tenth Edition.* Copyright © 2007 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved. Other data are adapted from Dunbar, Hoover, Frisbie, & Mengeling, *Iowa Tests of Basic Skills: 2005 Norms and Score Conversion, Complete and Core Batteries, Form K.* Copyright © 2008 by The University of Iowa. All rights reserved. Reproduced with permission of the Riverside Publishing Company.

hypothetical students on the mathematics subtest of two published tests if this view is adopted.

The students have these characteristics: Student A is one year behind in terms of grade equivalents, Student B is at grade level, and Student C is one year ahead. Each year, the students' grade equivalents show a one-year "growth" over the preceding year. But look at the percentile ranks corresponding to their scores: Student A, who starts out one year behind, has to *exceed more persons* in the norm group to maintain a one-year-behind grade equivalent. Being one year behind in second grade means being at the 12th or 25th percentile. However, one year behind in Grade 8 means being around the 39th or 43rd percentile. One has to move from the bottom of the group toward the middle. An opposite phenomenon occurs for Student C, who begins one grade ahead at around the 86th or 93rd percentile. In this case, the student can fall behind more and more students and still be "one year ahead." Students who are at grade level (Student B) have raw scores equivalent to the average. By definition, the average at a grade is assigned one year's growth from the preceding year. Thus, only students who are exactly at the average each year will maintain their approximate percentile rank from year to year.

An alternate norm-referenced definition is the *percentile view of normal growth*: A student shows normal growth if that student maintains the same position (i.e., percentile rank) in the norm from year to year. Figure 16.11 shows examples of what happens to a student's grade-equivalent score if that student's *percentile rank stays the same each year*.

Lower-scoring students (such as Students A and C)—even though they do not change their position in the norm group—have grade equivalents indicating they are further and further behind. An opposite trend occurs for initially high-scoring students. The exact magnitude of this falling-behind phenomenon will vary from one publisher's test to another's and depends on the student's percentile rank. The grade-equivalent scales of some tests are created to minimize the falling-behind effect. Students close to the 50th percentile will exhibit less of the falling-behind effect than will those further from the center of the distribution. The reasons for this effect are twofold: (1) the line connecting the medians of the distributions at each grade level tends to flatten out at higher grades rather than being a diagonal line (that is, the median gain decreases as grade increases), and (2) scores at upper grades become more spread out, spanning a larger range than scores at lower grades.

Unequal Units The grade-equivalent score scale does not have a one-to-one correspondence with the number of questions a student answers correctly on a test. This means, for example, that

FIGURE 16.11 **Examples of changes in the grade-equivalent score for four hypothetical students as each student's percentile rank remains the same from second through eighth grade.**

	Stanford Achievement Tests, **Total Mathematics**				***Iowa Tests of Basic Skills,*** **Total Mathematics**			
	Student A: "Below grade level" (*PR* = 16 each year)		**Student B: "Above grade level" (*P R* = 84 each year)**		**Student C: "Below grade level" (*P R* = 16 each year)**		**Student D: "Above grade level" (*PR* = 84 each year)**	
Grade placement at the time of testing	***GE***	**"Grades behind"**	***GE***	**"Grades ahead"**	***GE***	**"Grades behind"**	***GE***	**"Grades ahead"**
3.3	2.0	1.3	5.1	1.8	2.2	1.1	4.2	0.9
4.3	2.6	1.7	6.8	2.5	3.0	1.3	5.5	1.2
5.3	3.1	2.2	8.6	3.3	3.6	1.7	6.9	1.6
6.3	3.6	2.7	10.2	3.9	4.3	2.0	8.6	2.3
7.3	4.6	2.7	above 12.9		4.9	2.4	10.1	2.8
8.3	5.3	3.0	above 12.9		5.5	2.8	12.4	4.1

Source: Data reproduced from the *2007 Fall Supplemental Multilevel Norms Book: Stanford Achievement Test: Tenth Edition.* Copyright © 2007 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved. Other data are adapted from Dunbar, Hoover, Frisbie, & Mengeling, *Iowa Tests of Basic Skills: 2005 Norms and Score Conversion, Complete and Core Batteries, Form K.* Copyright © 2008 by The University of Iowa. All rights reserved. Reproduced with permission of the Riverside Publishing Company.

students in the middle of the distribution who get one more item correct are likely to raise their grade-equivalent scores by only one tenth (i.e., one "month"). For students in the upper part of the distribution, however, one additional correct item may result in an increment of several tenths (several "months" of growth). As a result of these unequal units, calculating averages using grade equivalents becomes problematic.

Grade Mean Equivalents

Because it is problematic to average grade-equivalent scores, some publishers (e.g., CTB/McGraw-Hill) have tried other ways to give schools information on how well their students performed on the average. One technique is to report the **grade mean equivalent** that tells the grade placement of a group's average extended scale score. Instead of averaging grade-equivalent scores directly, you first average the extended scale scores. Second, you look up the grade-equivalent that corresponds to this average extended scale score (CTB/McGraw-Hill, 2008). This averaging of extended scores is also problematic, however. There is evidence to suggest that extended score scales do not have equal units of measurement either, even though some test developers claim they do (Hoover, 1984a, 1984b). If this is the case, then their averages and the grade mean equivalents on which they are based would be just as problematic as averaging grade equivalents.

Recommendations

In light of the problems with grade equivalents, you may wonder why they are used at all. Indeed, many assessment specialists believe they should be eliminated. Yet such scores are popular with teachers and administrators who are generally unaware of the complex criticisms. Teachers and school administrators have a real need for at least some crude measure of educational development or growth that they can relate to years of schooling. Despite the technical difficulties in doing so, grade equivalents seem intuitively to be a "natural metric." Some assessment specialists recommend extended standard scores as measures of growth, but they possess many of the same interpretive problems as grade equivalents, and because they cannot be easily referenced to grade levels, their interpretation can be confusing.

You should use grade-equivalent scores as coarse indicators of educational development or growth but do so only when you report them with their corresponding percentile ranks. Grade equivalents are norm-referenced growth indicators. If you want information about the content of a student's learning, you need to look carefully at the kinds of performances the student can do. To do that, you need to review for each student the kinds of test items the student answered correctly. When you do this, of course, you are making criterion-referenced interpretations.

Summary of Grade Equivalents

As a summary of grade-equivalent scores, consider the situation in which a school administered a published, norm-referenced achievement test to third graders in May. Further, assume that the school's teachers have judged the assessment to match the curriculum validly and to be an appropriate way to assess the students. Finally, assume that the publisher's norms are appropriate. Then, even with all these nice assumptions, each of the following statements—except the first—are *false*:

1. Pat's Reading subtest grade-equivalent score is 3.8. This suggests that she is an average third-grade reader.
2. Ramon's Arithmetic subtest grade equivalent is 4.6. This means that he knows arithmetic as well as the typical fourth grader who is at the end of the 6th month of school.
3. Melba's Arithmetic subtest grade equivalent is 6.7. This suggests that next year she ought to take arithmetic with the sixth graders.
4. Debbie's Reading subtest grade equivalent is 2.3. This means she has mastered three tenths of the second-grade reading skills.
5. John's grade-equivalent profile is Vocabulary = 6.2, Reading = 7.1, Language = 7.1, Work-Study Skills = 7.2, Arithmetic = 6.7. This means that his weak areas are vocabulary and arithmetic.
6. Two of Sally's grade equivalents are Language = 4.5 and Arithmetic = 4.5. Because her language and arithmetic grade equivalents are the same, we conclude that her language and arithmetic ability are about equal.
7. Half of this school's second graders have grade equivalents below grade level. This means that instructional quality is generally poor.
8. This year one teacher was assigned all of the students whose assessment scores were in the bottom three stanines. The average of her class's grade equivalents this May was further below grade level than the class's average last year. This means that her instruction has been ineffective for the class as a whole.

GENERAL GUIDELINES FOR SCORE INTERPRETATION

Figure 16.12 summarizes the various norm-referenced scores discussed in this chapter. Although each type of score describes the student's location in a norm group, each does so differently. The easiest type of score to explain to parents and students is a percentile rank. Various types of linear standard scores require an understanding of the mean and standard deviation for their meaning to become clear. Usually you will need to interpret normalized standard scores in conjunction with percentile ranks. From test to test, normalized standard scores will have the same percentage of cases associated with them. Consequently, their meaning remains fairly constant as long as a normal distribution can be assumed. Grade equivalents and extended standard scores provide scores along an educational growth continuum, but because of their inherent technical complexities, teachers and school administrators may misinterpret them. Limit using grade equivalents to gross estimates of yearly student growth. Use them only when you accompany them with percentile ranks. Use percentile ranks to compare an individual student's performance in different curriculum areas.

Teachers and school administrators should consider the following points when interpreting student scores on norm-referenced standardized tests:

1. *Look for unexpected patterns of scores.* An assessment should confirm what a teacher knows from daily interactions with a student; unusually high or low scores for a student should be a signal for exploring instructional implications.
2. *Seek an explanation for patterns.* Ask why a student is higher in one subject than another. Check for motivation, special interests, special difficulties, and so on.
3. *Don't expect surprises for every student.* Most students' assessment results should be as you expect from their performance in class. A valid assessment should confirm your observations.
4. *Small differences in subtest scores should be viewed as chance fluctuations.* Use the standard error of measurement (Chapter 4) to help decide whether differences are large enough to have instructional significance.
5. *Use information from various assessments and observation to explain performance on other assessments.* Students low in reading comprehension may perform poorly on the social studies subtest, for example.

You may wish to try your hand at implementing these general guidelines by reviewing the case

FIGURE 16.12 **How to interpret different types of norm-referenced scores.**

Type of score	Interpretation	Score	Examples of interpretations
Percentile rank linear standard score (*z*-score)	Percentage of scores in a distribution below this point.	*PR* = 60	"60% of the raw scores are lower than this score."
	Number of standard deviation units a score is above (or below) the mean of a given distribution.	*z* = +1.5	"This raw score is located 1.5 standard deviations *above* the mean."
		z = −1.2	"This raw score is located 1.2 standard deviations *below* the mean."
Linear standard score (*SS*-score or 50± 10 system)	Location of score in a distribution having a mean of 50 and a standard deviation of 10. (Note: For other systems, substitute in these statements that system's mean and standard deviation.)	*SS* = 65	"This raw score is located 1.5 standard deviations *above* the mean in a distribution whose mean is 50 and whose standard deviation is 10."
		SS = 38	"This raw score is located 1.2 standard deviations *below* the mean in a distribution whose mean is 50 and whose standard deviation is 10."
Stanine	Location of a score in a specific segment of a normal distribution of scores.	Stanine = 5	"This raw score is located in the middle 20% of a normal distribution of scores."
		Stanine = 9	"This raw score is located in the top 4% of a normal distribution of scores."
Normalized standard score (*T*-score or normalized 50± 10 system)	Location of score in a normal distribution having a mean of 50 and a standard deviation of 10. (Note: For other systems, substitute in these statements that system's mean and standard deviation [e.g., *DIQs* have a mean of 100 and a standard deviation of 16: This is a 100± 16 system].)	*T* = 65	"This raw score is located 1.5 standard deviations above the mean in a normal distribution whose mean is 50 and whose standard deviation is 10. This score has a percentile rank of 84."
		T = 38	"This raw score is located 1.2 standard deviations below the mean in a normal distribution whose mean is 50 and whose standard deviation is 10. This score has a percentile rank of 12."
Extended standard score	Location of a score on an arbitrary scale of numbers that is anchored to some reference group.		(No interpretation is offered here because the systems are so arbitrary and unalike.)
Grade-equivalent score	The grade placement at which the raw score is average.	*GE* = 4.5	"This raw score is the obtained or estimated average for all pupils whose grade placement is at the 5th month of the fourth grade."

Source: Adapted from *Measuring Student Achievement and Aptitude* (2nd ed., p. 99), by C. M. Lindvall and A. J. Nitko, 1975, New York: Harcourt Brace Jovanovich. Reproduced by permission of the authors.

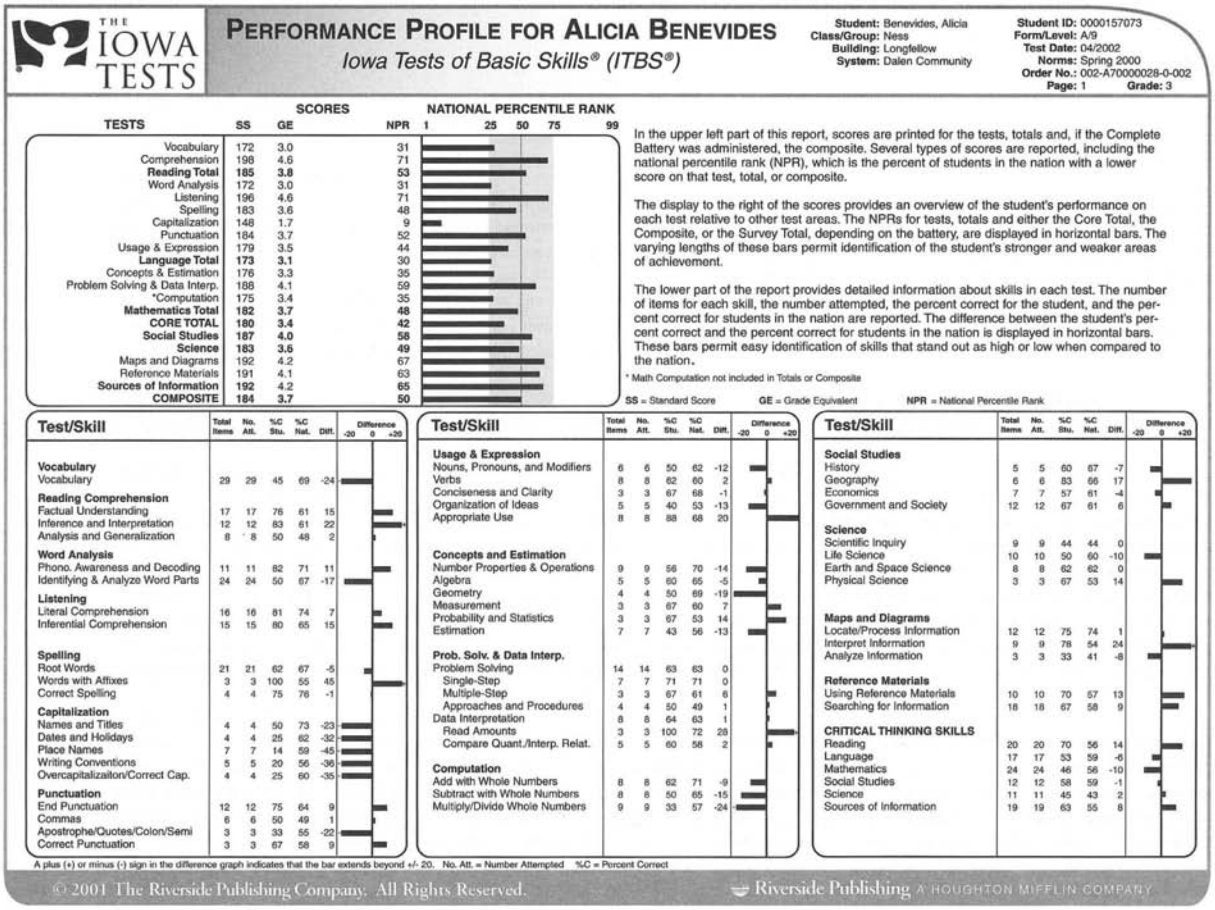
presented in Figure 16.13. Your interpretation of Alicia Benevides's report may be different from that of the computer.

Types of Questions Parents Ask and Suggested Ways of Answering Them

A teacher has the most direct contact with parents regarding norm-referenced score reports. Parents call the teacher first if they have questions about students' standardized test results. You must be prepared, therefore, to explain students' test results to their parents. Studying the concepts and principles in this chapter is a prerequisite for effectively communicating to parents.

Figure 16.14 contains examples of many of the questions parents ask when they receive

FIGURE 16.13 Example of an individual performance profile report.



THE IOWA TESTS

PERFORMANCE PROFILE FOR ALICIA BENEVIDES

Iowa Tests of Basic Skills® (ITBS®)

Student: Benevides, Alicia
Class/Group: Ness
Building: Longfellow
System: Dalen Community

Student ID: 0000157073
Form/Level: A/9
Test Date: 04/2002
Norms: Spring 2000
Order No.: 002-A70000028-0-002
Page: 1 **Grade:** 3

TESTS	SCORES SS	 GE	 NPR	NATIONAL PERCENTILE RANK 1 25 50 75 99
Vocabulary	172	3.0	31	
Comprehension	198	4.6	71	
Reading Total	**185**	**3.8**	**53**	
Word Analysis	172	3.0	31	
Listening	196	4.6	71	
Spelling	183	3.6	48	
Capitalization	148	1.7	9	
Punctuation	184	3.7	52	
Usage & Expression	179	3.5	44	
Language Total	**173**	**3.1**	**30**	
Concepts & Estimation	176	3.3	35	
Problem Solving & Data Interp.	188	4.1	59	
*Computation	175	3.4	35	
Mathematics Total	**182**	**3.7**	**48**	
CORE TOTAL	**180**	**3.4**	**42**	
Social Studies	**187**	**4.0**	**58**	
Science	**183**	**3.6**	**49**	
Maps and Diagrams	192	4.2	67	
Reference Materials	191	4.1	63	
Sources of Information	**192**	**4.2**	**65**	
COMPOSITE	**184**	**3.7**	**50**	

In the upper left part of this report, scores are printed for the tests, totals and, if the Complete Battery was administered, the composite. Several types of scores are reported, including the national percentile rank (NPR), which is the percent of students in the nation with a lower score on that test, total, or composite.

The display to the right of the scores provides an overview of the student's performance on each test relative to other test areas. The NPRs for tests, totals and either the Core Total, the Composite, or the Survey Total, depending on the battery, are displayed in horizontal bars. The varying lengths of these bars permit identification of the student's stronger and weaker areas of achievement.

The lower part of the report provides detailed information about skills in each test. The number of items for each skill, the number attempted, the percent correct for the student, and the percent correct for students in the nation are reported. The difference between the student's percent correct and the percent correct for students in the nation is displayed in horizontal bars. These bars permit easy identification of skills that stand out as high or low when compared to the nation.

* Math Computation not included in Totals or Composite

SS = Standard Score **GE** = Grade Equivalent **NPR** = National Percentile Rank

Test/Skill	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.	Difference -20 0 +20
Vocabulary						
Vocabulary	29	29	45	69	-24	
Reading Comprehension						
Factual Understanding	17	17	76	61	15	
Inference and Interpretation	12	12	83	61	22	
Analysis and Generalization	8	8	50	48	2	
Word Analysis						
Phono. Awareness and Decoding	11	11	82	71	11	
Identifying & Analyze Word Parts	24	24	50	67	-17	
Listening						
Literal Comprehension	16	16	81	74	7	
Inferential Comprehension	15	15	80	65	15	
Spelling						
Root Words	21	21	62	67	-5	
Words with Affixes	3	3	100	55	45	
Correct Spelling	4	4	75	76	-1	
Capitalization						
Names and Titles	4	4	50	73	-23	
Dates and Holidays	4	4	25	62	-32	
Place Names	7	7	14	59	-45	
Writing Conventions	5	5	20	56	-36	
Overcapitalizaiton/Correct Cap.	4	4	25	60	-35	
Punctuation						
End Punctuation	12	12	75	64	9	
Commas	6	6	50	49	1	
Apostrophe/Quotes/Colon/Semi	3	3	33	55	-22	
Correct Punctuation	3	3	67	58	9	

Test/Skill	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.	Difference -20 0 +20
Usage & Expression						
Nouns, Pronouns, and Modifiers	6	6	50	62	-12	
Verbs	8	8	62	60	2	
Conciseness and Clarity	3	3	67	68	-1	
Organization of Ideas	5	5	40	53	-13	
Appropriate Use	8	8	88	68	20	
Concepts and Estimation						
Number Properties & Operations	9	9	56	70	-14	
Algebra	5	5	60	65	-5	
Geometry	4	4	50	69	-19	
Measurement	3	3	67	60	7	
Probability and Statistics	3	3	67	53	14	
Estimation	7	7	43	56	-13	
Prob. Solv. & Data Interp.						
Problem Solving	14	14	63	63	0	
Single-Step	7	7	71	71	0	
Multiple-Step	3	3	67	61	6	
Approaches and Procedures	4	4	50	49	1	
Data Interpretation	8	8	64	63	1	
Read Amounts	3	3	100	72	28	
Compare Quant./Interp. Relat.	5	5	60	58	2	
Computation						
Add with Whole Numbers	8	8	62	71	-9	
Subtract with Whole Numbers	8	8	50	65	-15	
Multiply/Divide Whole Numbers	9	9	33	57	-24	

Test/Skill	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.	Difference -20 0 +20
Social Studies						
History	5	5	60	67	-7	
Geography	6	6	83	66	17	
Economics	7	7	57	61	-4	
Government and Society	12	12	67	61	6	
Science						
Scientific Inquiry	9	9	44	44	0	
Life Science	10	10	50	60	-10	
Earth and Space Science	8	8	62	62	0	
Physical Science	3	3	67	53	14	
Maps and Diagrams						
Locate/Process Information	12	12	75	74	1	
Interpret Information	9	9	78	54	24	
Analyze Information	3	3	33	41	-8	
Reference Materials						
Using Reference Materials	10	10	70	57	13	
Searching for Information	18	18	67	58	9	
CRITICAL THINKING SKILLS						
Reading	20	20	70	56	14	
Language	17	17	53	59	-6	
Mathematics	24	24	46	56	-10	
Social Studies	12	12	58	59	-1	
Science	11	11	45	43	2	
Sources of Information	19	19	63	55	8	

A plus (+) or minus (-) sign in the difference graph indicates that the bar extends beyond +/- 20. No. Att. = Number Attempted %C = Percent Correct

Riverside Publishing A HOUGHTON MIFFLIN COMPANY

Source: S. Dunbar, H. D. Hoover, D. A. Frisbie, and K. R. Oberley, 2008. Copyright © 2008 by The University of Iowa. All rights reserved. Reproduced from the *Iowa Tests®, Interpretive Guide for School Administrators,* p. 114, with permission of the Riverside Publishing Company.

standardized test results from a school. The questions are organized into five categories: standing, growth, improvement needed, strengths, and intelligence. You should be prepared to answer questions in these categories. The table contains suggestions for answering each category of questions. Note that we indicate which type of norm-referenced score to use. Although other scores might be used, we believe the ones suggested will be most helpful to your explanation. Notice, too, that we suggest always using a student's classroom performance to complement and explain the student's standardized test results. Because in the majority of cases students' standardized test performance will be quite consistent with their classroom performance, using students' classroom performance to illustrate their standardized test performance will help you reinforce to the parents your assessment of the students.

Parent Misunderstandings

Parents also have misunderstandings about what norm-referenced test scores mean. We have already discussed many of the misconceptions and limitations in this chapter. The following list *summarizes common parent misunderstandings* that you need to be clear about before you can help parents correct them:

1. The grade-equivalent score tells which grade the student should be in.

FIGURE 16.14 **How to answer parents' questions about standardized test results.**

Category	Examples of questions	Suggestions for answering
Standing	■ How is my child doing compared to others? ■ Is my child's progress normal for his or her grade?	Use percentile ranks to describe standing. Explain that a standardized test gives partial information only. Use information from classroom performance to explain progress.
Growth	■ Has my child's growth been as much as it should be?	Use grade-equivalent scores to show progress from previous years. Use composite scores (i.e., all subjects combined) to show general growth; use scores from each subject to explain growth in particular curricular areas. Obtain past performance information from the child's cumulative folder. Use information from classroom performance to explain growth.
Improvement needed	■ Does my child have any learning weaknesses? ■ How can I help improve my child's learning?	Use percentile ranks to identify relative weaknesses. Use information about a student's performance to clusters of similar questions to pinpoint weaknesses. Use information from class performance to explain specific weaknesses. Don't overemphasize weaknesses. Explain a student's relative strengths, too; give specific suggestions as to how parents can help.
Strengths	■ What does my child do well?	Use percentile ranks to pick out areas of relative strengths. Use class information to illustrate the point. Make suggestions for how parents can help improve these areas even more.
Intelligence	■ How smart is my child? Is my child gifted?	Explain that an achievement test is not an intelligence test. Explain that an achievement test is very sensitive to what was taught in class and that high scores may only reflect specific opportunities to learn. Use class information to illustrate your points.

Source: This figure is based on suggestions in Hoover et al., 1993.

2. The percentile rank and percent-correct scores mean the same thing.
3. The percentile rank norm group consists of only the students in a particular classroom.
4. "Average" is the standard to beat.
5. Small changes in percentile ranks over time are meaningful.
6. Percent-correct scores below 70 are failing.
7. If you get a perfect score, your percentile rank must be 99 (Hoover et al., 1993b, pp. 103–105).

CONCLUSION

We hope this chapter has given you enough detail about the most common norm-referenced scores that you can use them thoughtfully in making decisions about your students, class, and school. We hope, too, that this chapter has given you enough information about norm-referenced scores to communicate their meaning to students, parents, school board members, and other interested community members. A central point is that no matter which norm-referenced score you are using, the score derives its meaning from comparisons to other test takers in a norm group, so it is important to know the nature of the norm group and how it is relevant to your purpose.

The next chapter presents information about finding and evaluating published tests. The next chapter also describes briefly how standardized tests are developed.

EXERCISES

1. A student takes a test during the middle of the school year. By mistake, the student's teacher uses the norms tables published for the end of the school year to look up the student's percentile rank. What effect does this error have on the percentile ranks the teacher reports? What would be the effect in this case if the teacher used the norms tables from the beginning of the year?

2. Read each of these statements and decide to which norm-referenced score(s) each mainly refers. Justify your choice(s) to your classmates.
   a. In this skewed distribution, John's score places him one standard deviation below the mean.
   b. Roberto's test score is the same as the average score of students tested in the 4th month of fifth grade.
   c. Because Bill's score increased this year, I know that his general educational development has increased, even though his position in the norm group remained the same.
   d. Nancy's score is 5 because it is located in the middle 20% of a normal distribution.
3. Judge each of the following statements true or false. Explain the basis for your judgment in each case.
   a. A person's percentile rank is 45. This means that the person's raw score was the same as 45% of the group assessed.
   b. Kaiko's arithmetic assessment score is 40. The class's mean score is 45, and its standard deviation is 10. Therefore, Kaiko is located one standard deviation below the mean.
   c. The norms tables show that the distribution of deviation IQ scores on a school ability test is approximately normal in form. This means that for the people in the norm groups, the intellectual ability that naturally underlies the scores is normally distributed.
4. Figure 16.15 shows several types of normalized scores. Use the relationships between the scores to complete the table and thereby show how various scores are related to one another. The first two are completed for you. You may use Figure 16.6 for assistance.

FIGURE 16.15 Use with Exercise 4.

Percentile rank	Stanine	z_n	**DIQ** (*SD* = 15)	*T*-score
99.9	9	+3.00		
98				
84				
50				
16				
2				
0.1				

5. Figure 16.16 shows part of a norms table that might appear in a manual of a standardized achievement test. The table shows selected raw scores, grade-equivalent scores, and percentile ranks for the publisher's standardization sample (i.e., norm group). Assume that (a) the local school system has judged the test's content to be a good match to its curriculum, (b) the norm data were collected during the 7th month of the fourth grade, (c) the norms are appropriate for use with the local school system, (d) the publisher has computed grade equivalents and percentile ranks in the usual way and with no errors, and (e) the school tested the students in April.

   Use the table and your knowledge of norm-referenced frameworks to judge each of the following statements as true or false. Explain and justify your position in each case.
   a. James is a fourth-grade student with a grade-equivalent profile of $V = 6.2$, $R = 5.6$, $L = 5.6$, $W = 5.6$, $A = 6.2$. Decide whether each of the following conclusions is true or false, and explain the basis for your judgment.
      i. James should be in fifth grade.
      ii. James is strongest in vocabulary and arithmetic.
      iii. James's scores are above average for his grade.
   b. Fourth grader Jasmine's raw score on reading is 50, and on language it is 30. Decide whether each of the following conclusions is true or false, and explain your decision.
      i. Jasmine is more able in reading because her raw score in reading is higher.
      ii. Because Jasmine's grade-equivalent scores are equal, she is equally able in reading and vocabulary (relative to the norm group).
      iii. Jasmine is more able in language than reading (relative to the norm group) because her percentile rank in language is higher.

FIGURE 16.16 Use with Exercise 5.

	Vocabulary (*V*)		**Reading (*R*)**		**Language (*L*)**		**Work-study (*W*)**		**Arithmetic (*A*)**	
Raw score	***GE***	***PR***	***GE***	***PR***	***GE***	***PR***	***GE***	***PR***	***GE***	***PR***
5	1.8	1	1.6	1	1.9	1	2.3	1	2.5	1
20	4.1	34	3.3	17	4.4	41	5.6	74	5.5	65
30	5.1	61	4.2	36	5.6	75	7.0	96	6.2	74
40	6.2	74	4.8	52	6.4	86	7.6	99	6.9	97
50	7.0	96	5.6	74	7.9	99	8.0	99	7.7	99
70			8.1	99						

Finding and Evaluating Published Assessments

KEY CONCEPTS

1. You can search printed materials, online resources, and personal contacts to locate information about published assessments.
2. After locating published tests, obtain and read evaluations or reviews of these tests.
3. Internet searches and test publishers' Websites will help you locate computerized testing materials.
4. There are also print and online resources for locating unpublished test materials, which can be useful in research and evaluation projects.
5. Some publishers restrict the sale of test materials.
6. To evaluate and select a test or assessment, clarify your purpose, obtain and review assessment materials, and try them out in a pilot study.
7. Standardized test development should follow a set of steps.

IMPORTANT TERMS

ETS Test Collection
external assessment procedure
*Mental Measurements Yearbook*s (*MMY*s)
specimen set
Standards for Educational and Psychological Testing
technical manual
Test Critiques
Tests in Microfiche
Tests in Print
universal design

From Chapter 17 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

LOCATING A PUBLISHED TEST

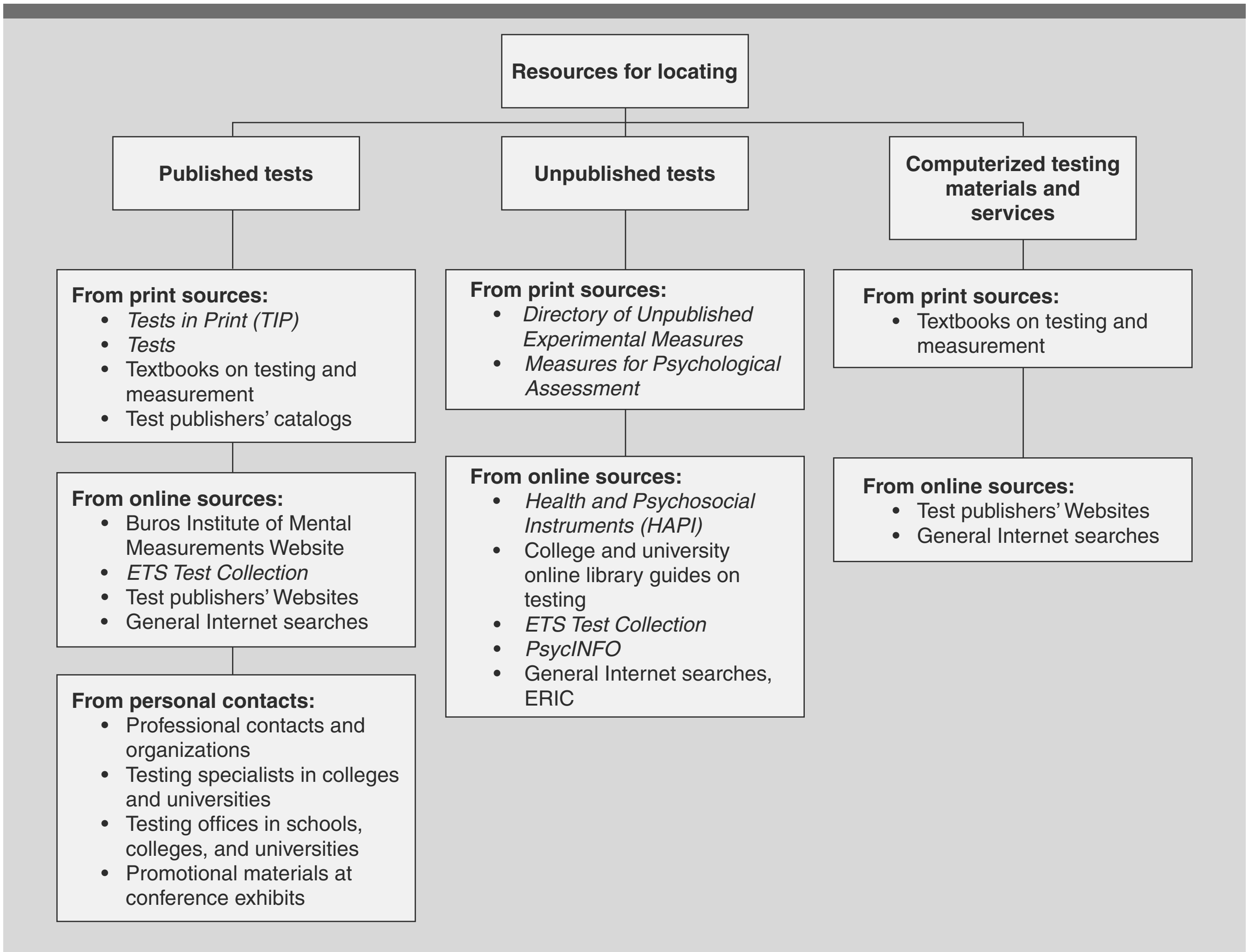
Suppose you want to locate a test that you can use to diagnose students' reading problems, assess students' self-concept, or assess some other category of students' characteristics. How do you locate possible tests for such purposes if you do not even know the name(s) of the particular test(s)? You can begin your search in one of three ways: (a) search printed materials, (b) search online resources, or (c) make personal contacts with persons who may know what you are looking for. Figure 17.1 is an overview of the available sources for locating a test. We describe each of these resources in this section.

Locating Published Tests From Print Sources

Four print resources are available. Three of these resources are likely to be in your library: *Tests in Print, Tests,* and textbooks on testing and measurement. You are not likely to find the fourth, test publishers' catalogs, in a library.

Tests in Print (TIP) VII The **Tests in Print** (7th ed.; Murphy, Spies, & Plake, 2006) is a test bibliography that contains information on more than 4,000 commercially available instruments. You can use this resource to identify appropriate tests, locate reviews of tests in the *Mental Measurements Yearbook*s (discussed later), and find publishers' addresses. To appear as an entry in *Tests in Print,* a test must be currently in print and must be published in English. Each entry includes the following information: a description of the test and its purpose, information on population and scoring, test editions available and their price, name of the publisher, and location of the test's review in the *Mental Measurements Yearbook*. (See Figure 17.4 for an example of a *TIP*

FIGURE 17.1 **Where to look to locate an assessment tool.**


Resources for locating
Published tests
Unpublished tests
Computerized testing materials and services
From print sources:
• *Tests in Print (TIP)*
• *Tests*
• Textbooks on testing and measurement
• Test publishers' catalogs
From online sources:
• Buros Institute of Mental Measurements Website
• *ETS Test Collection*
• Test publishers' Websites
• General Internet searches
From personal contacts:
• Professional contacts and organizations
• Testing specialists in colleges and universities
• Testing offices in schools, colleges, and universities
• Promotional materials at conference exhibits
From print sources:
• *Directory of Unpublished Experimental Measures*
• *Measures for Psychological Assessment*
From online sources:
• *Health and Psychosocial Instruments (HAPI)*
• College and university online library guides on testing
• *ETS Test Collection*
• *PsycINFO*
• General Internet searches, ERIC
From print sources:
• Textbooks on testing and measurement
From online sources:
• Test publishers' Websites
• General Internet searches

entry.) The *TIP VII* also provides listing notations on out-of-print tests.

You locate a test by using one of the book's five indexes:

1. If you know or have some idea of the test's name, look in the Index of Titles. The index lists all of the tests in *TIP VII* plus all those that are out of print.
2. If you do not know the test name, but know the category or type of test, look in the Classified Subject Index. All tests in the book are grouped into categories (e.g., Social Studies, Speech & Hearing, etc.), with the individual tests listed alphabetically under the category.
3. If you know the name of the test author or person who has reviewed the test in one of the *Mental Measurements Yearbook*s, look in the Index of Names.
4. If you know the type of score a test may yield, look in the Score Index. For example, you may recall from your reading that a test contained an "aggression/hostility" score or an "enjoyment of mathematics" score, but do not know the names of the respective tests. The Score Index lists all such scores alphabetically for the tests included in *TIP VII*. The entries in this index are very specific to the tests that provide the scores. This means that your definition of the score may differ from a test publisher's or that different publishers may score the same student trait under different names. Thus, check all alternative or related score names before concluding that a test is not included in *TIP VII*.
5. If you know the acronym for a test but not its complete name, look in the Index of Acronyms. You may want to use this index, for example, if you recall that there is a test called the *DAT* that was used in counseling and want to locate it. The Index of Acronyms lists two *DAT*s: *Dental Admissions Test* and *Differential Aptitude Tests*. Because you want the counseling test, it is likely the latter rather than the former.

***Tests: A Comprehensive Reference for Assessments in Psychology, Education, and Business* (6th ed.)** This reference (Maddox, 2008) lists and describes approximately 2,000 tests, but gives no evaluations of them. (The test evaluations are given in *Test Critiques*, a companion volume, described later.) *Tests* is divided into three primary groups—psychology, education, and business—and 90 subcategories. Each listing describes the test and its purpose, for whom it is intended, scoring procedures, costs, and publisher.

The organization of the indexes makes locating tests easy. *Tests* has several indexes: title, publisher, computer-scoring, hearing impaired, visually impaired, physically impaired, out of print, tests found in the fifth but not in the sixth edition, and foreign language availability. Publishers' Websites are also listed.

Textbooks on Testing and Measurement A number of textbooks list, describe, and (sometimes) review selected tests. If you are looking for a test in a specific area, looking in the index of a textbook in the area may be a useful way to see which tests are frequently used. (Appendix K in this book lists a selection of published tests in several areas.) A textbook, however, is not a comprehensive source for information about tests because (a) tests are often selected for inclusion primarily for their merits in illustrating an author's point, (b) space permits only a few tests being mentioned, (c) often only the most popular or easily available tests are mentioned or illustrated, and (d) no single author is aware of all available tests.

Test Publishers' Catalogs An important way to get information about a test is directly from the test publisher. (See Appendix L for a partial list of publishers and their Websites.) Most test publishers have catalogs that describe the tests they publish in detail. A publisher's catalog is especially helpful for finding out about current editions of tests along with information about scoring services, costs, and how to obtain specimen sets, test manuals, and technical reports. Current information of this sort is seldom found in other print sources. Your school's testing office and the testing and measurement office of a college or university usually maintain collections of recent catalogs.

Locating Published Tests Online

Buros Test Locator The home page of the Buros Institute of Mental Measurements (http://www.unl.edu/buros) may be navigated to locate its database of published tests. This site lists more than 4,000 commercially available tests.

ETS Test Collection The **ETS Test Collection** is a database of approximately 20,000 tests and other assessment instruments, both published and unpublished. Some of the instruments listed in the database are out of print, some are available from publishers, some are available from the test authors, and some can be purchased and downloaded from ETS. Some of the tests are even from outside the United States. You may search for a test type, an author, or a title. Click the "Find a Test" tab from the Test Collection home page for further assistance in searching. You can access the *ETS Test Collection* database directly at http://www.ets.org/testcoll.

Test Publishers' Websites As we discussed earlier, an important way to get information about a test is directly from the test publisher. (See Appendix L for a partial list of publishers and their Websites.) Most test publishers have Websites that describe the tests they publish in detail. A publisher's Website is especially helpful for finding out about current editions of tests along with information about scoring services, costs, and how to obtain specimen sets, test manuals, and technical reports. Current information of this sort is seldom found in other print sources. Remember, however, that publishers' Websites are marketing tools, not objective sources of test information.

General Internet Searches If you are unable to locate a test through one of these online sources, you could try searching the Internet with the test title, author's name, or subject area. Including "test" or "assessment" with the subject-area key word helps to narrow the search. Usually, searching through http://www.eric.ed.gov will yield more relevant hits for educational tests than a general search, say on Google. You may also want to try searching on *PsycINFO* in your library.

FIGURE 17.2 **Examples of organizations that provide information on educational assessment.**

Professional associations prepare periodicals and other publications related to educational assessment, work toward improved assessment usage, and may be contacted to identify members who are experts in certain areas of educational assessment.

1. American Educational Research Association (Washington, DC) [http://www.aera.net]
2. Association for Assessment in Counseling and Education (Arlington,VA) [http://aac.ncat.edu]
3. International Reading Association (IRA) (Newark, DE) [http://www.reading.org]
4. National Association of Test Directors (NATD) [http://www.natd.org]
5. National Council on Measurement in Education (NCME) (Washington, DC) [http://www.ncme.org]

Educational Research Information Center (ERIC). Online services are provided.

1. Search ERIC [http://www.eric.ed.gov]

Research centers and regional laboratories invest in research on technical or policy issues in educational assessment. They have catalogs of these publications and sometimes answer inquiries about specific assessment issues.

1. Buros Institute of Mental Measurements (University of Nebraska) [http://www.unl.edu/buros]
2. Center for the Study of Testing, Evaluation, and Educational Policy (Boston College) [http://www.bc.edu/research/csteep]
3. Center for Research on Evaluations, Standards, and Student Testing (CRESST) (UCLA) [http://www.cse.ucla.edu]
4. Northwest Regional Educational Laboratory (NWREL) (Portland, OR) [http://www.nwrel.org]
5. Comprehensive Regional Assistance Centers [www.ccnetwork.org/where.html]
6. Mid-continent Research for Education and Learning (Mcrel)(Aurora, CO) [http://www.mcrel.org]

Nonprofit testing corporations offer a wide range of assessment services, conduct assessment research, and disseminate assessment information.

1. American College Testing Program (ACT) (Iowa City, IA) [http://www.act.org]
2. Educational Testing Service (ETS) (Princeton, NJ) [http://www.ets.org]

Nonprofit advocacy and public interest groups research matters of legality, individual rights, and public policy related to assessment.

1. National Center for Fair and Open Testing (Fair Test) (Cambridge, MA) [http://www.fairtest.org]

Locating Published Tests Through Personal Contacts

Professional Contacts and Organizations Figure 17.2 lists organizations that may help you locate published tests. Larger testing companies and agencies usually have an information and/or advisory office to answer questions that you can reach via toll-free telephone numbers. Professional organizations, such as the National Council on Measurement in Education, can sometimes help by referring you to a member in your local area who can be of assistance. Some professional associations whose focus is not on assessment per se may have special interest groups that are interested in specific issues such as performance, critical thinking, or classroom assessment. In some areas, federally funded research and development centers and regional laboratories have technical assistance offices that can help with testing problems. In some states, county-based school agencies, state-related school agencies, or technical assistance centers are specially organized to offer assistance in reviewing and using tests.

Testing Specialists in Colleges and Universities Testing and measurement professors at colleges and universities usually work in departments of educational research, educational psychology, measurement and statistics, counseling and guidance, or psychology. Ask the department chairperson for recommendations of persons to call.

Testing Offices in Schools, Colleges, and Universities The director of your school testing office or the school psychologist is frequently a useful resource. Many larger colleges and universities have testing offices designed to help their faculties and students with testing problems, and such offices are usually available to answer questions from the public as well.

Promotional Exhibits at Conferences If you attend a professional conference you may go to the exhibit area. The exhibits will have books and instructional materials, and may also have exhibits by test publishers. The exhibitor may have a test you are seeking or may put you in contact with a sales representative in your area who has the information you seek.

LOCATING EVALUATIONS OF PUBLISHED TESTS

In the preceding section we discussed how to find a test. But finding a test is only one part of the information you need to evaluate a published test. You will also need to locate published reviews of the test, preferably by reviewers who are competent assessment specialists and who are not associated with the test publisher. In this section we discuss three places to look for reviews of the test you have located: (a) printed materials, (b) online resources, or (c) direct contacts with persons who may know about the test. Figure 17.3 gives an overview of the resources that are available for locating test reviews.

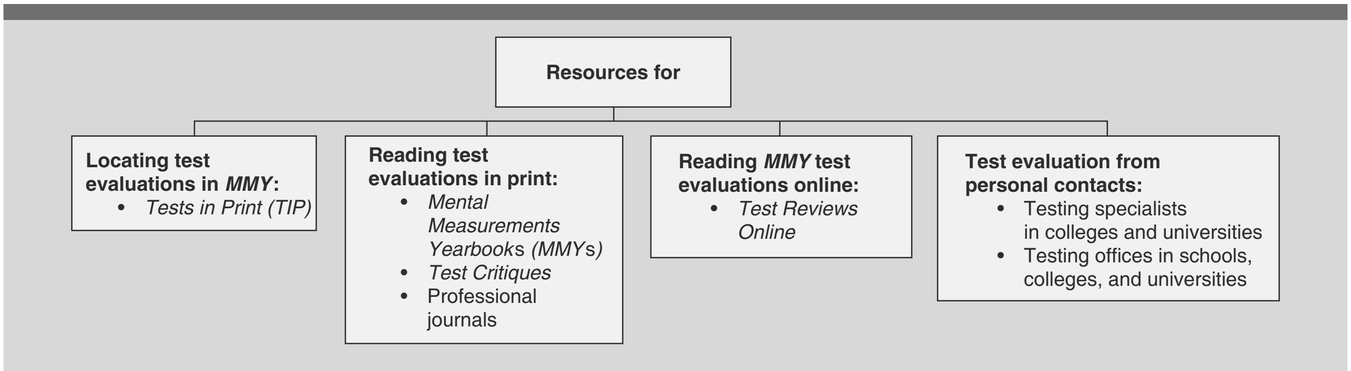
Locating Evaluations of Published Tests from Print Sources

Three print resources are likely to be in your academic library: *Mental Measurements Yearbook*s, *Test Critiques,* and professional journals.

Mental Measurements Yearbooks (MMYs) Among the most useful resources for locating information on tests are the publications of the Buros Institute of Mental Measurements (located at the University of Nebraska). The late Oscar K. Buros founded the institute and began a series of test bibliographies and **Mental Measurements Yearbooks (MMYs)**. The *Mental Measurements Yearbook*s (Buros, 1938 through present) are a series of volumes that critically evaluate many of the currently available published tests in English. Each volume supplements rather than replaces the earlier editions, so it is occasionally necessary to consult earlier volumes to obtain complete coverage of a test. One or more experts review each test especially for the *MMY*s, and each volume gives excerpted journal reviews as well. Each *Mental Measurements Yearbook* contains original reviews of hundreds of tests.

Each test entry is organized into five sections. The *description section* contains a test title, age or

FIGURE 17.3 **Where to locate published reviews of tests.**


Resources for
Locating test evaluations in *MMY*:
• *Tests in Print (TIP)*
Reading test evaluations in print:
• *Mental Measurements Yearbooks (MMYs)*
• *Test Critiques*
• Professional journals
Reading *MMY* test evaluations online:
• *Test Reviews Online*
Test evaluation from personal contacts:
• Testing specialists in colleges and universities
• Testing offices in schools, colleges, and universities

grade levels, publication dates, special comments, number and type of part scores, authors, publishers, references, and bibliographic information. Each entry is cross-referenced to previous reviews in earlier *MMY*s. The *development section* evaluates how well the publisher developed the test using professionally accepted standards, including the use of empirical data in the development process (recall our discussion of empirically developed test in Chapter 16). The *technical section* evaluates the tests standardization procedure, reliability, and validity. The *commentary section* contains the reviewer's overall evaluation of the test. The *summary section* is a concise wrap-up of the reviewer's opinion of the test. Names and addresses of hundreds of test publishers are listed in each *MMY*. A disadvantage of the printed *MMY*s is that because of the publication lag, editions of tests reviewed may not correspond to publishers' newest editions.

To locate a test evaluation, you need to know the *MMY* volume in which the test appears. If your test title appears in *Tests in Print VII*, that publication will tell you the review's *MMY* volume and entry number. You may also use the *Test Reviews Online* described in the last section by accessing it on the Internet through the Buros Institute home page. Figure 17.4 shows how a page in the *MMY* is laid out. This will give you a better sense of what to expect from this resource. For an online lesson in how to use an *MMY* test review see Nitko (2005b) at http://www.unl.edu/buros/bimm/html/lesson01.html.

Test Critiques (Volumes I–XI) **Test Critiques** (Keyser & Sweetland, 2005) is a series of volumes that reviews the most frequently used tests in business, education, and psychology. A testing specialist reviews each test. Entries cover an introduction to the test, practical uses and applications, technical aspects, and an overall evaluation of the test. You may use *Tests: A Comprehensive Reference for Assessments in Psychology, Education, and Business*, described earlier, to locate the *Test Critiques* volume in which the test is found.

Professional Journals Professional journals in a field often review tests that have potential application in a particular area, such as reading, mathematics, child development, or learning disabilities. Specialized testing and measurement journals review tests that have a wide appeal to school practitioners and psychologists. Among the journals that often report test reviews are *Developmental Medicine and Child Neurology; Journal of Educational Measurement; Journal of Learning Disabilities; Journal of Personality Assessment; Journal of Reading; Journal of School Psychology; Journal of Special Education; Measurement and Evaluation in Guidance; Modern Language Journal; Psychological Reports; Psychology in the Schools;* and *Reading Teacher*.

Bibliographic information about these and other journals (including those that review testing books) appears at the back of some *Mental Measurements Yearbook*s. Journal references are indexed in such sources as the *Education Index, Dissertation Abstracts, Research Studies in Education, Psychological Abstracts*, and *Research in Education* (ERIC).

Locating Evaluations of Published Tests Online

One source for obtaining test reviews online is *Test Reviews Online* from the Buros Institute (http://www.unl.edu/buros). Test reviews from *MMY*s 9 through 17 are available. More than 2,000 test reviews are available, and can be seen and downloaded for a fee. Your university library may have a subscription to *MMY* that allows its members to use this resource.

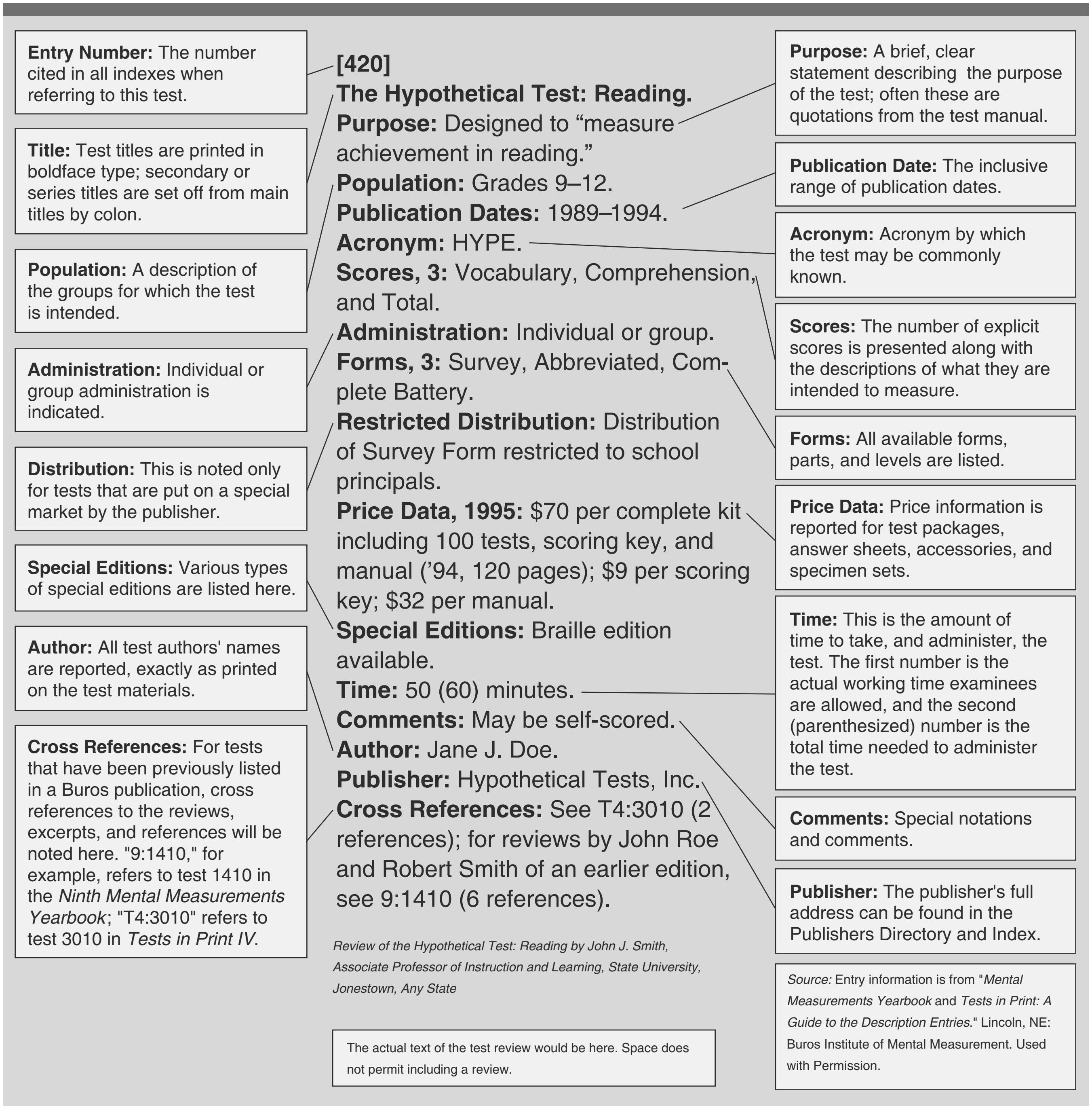
Locating Evaluations of Published Tests Through Personal Contacts

Testing Specialists in Colleges and Universities

Just as your professional contacts may help you locate a test, these same people may help you evaluate a test. Testing and measurement professors at colleges and universities usually work in departments of educational research, educational psychology, measurement and statistics, counseling and guidance, or psychology. Call the department chairperson for recommendations of persons to call. A testing specialist may have personal experience with a particular test and be willing to share that experience with you.

Testing Offices in Schools, Colleges, and Universities The director of your school testing office or the school psychologist may be a useful resource. Many larger colleges and universities have testing offices designed to help their faculties and students with testing problems. Such offices are usually available to answer questions from the public, as well.

FIGURE 17.4 **Layout of a *Mental Measurements Yearbook* review entry for a hypothetical test.**



Entry Number: The number cited in all indexes when referring to this test.

Title: Test titles are printed in boldface type; secondary or series titles are set off from main titles by colon.

Population: A description of the groups for which the test is intended.

Administration: Individual or group administration is indicated.

Distribution: This is noted only for tests that are put on a special market by the publisher.

Special Editions: Various types of special editions are listed here.

Author: All test authors' names are reported, exactly as printed on the test materials.

Cross References: For tests that have been previously listed in a Buros publication, cross references to the reviews, excerpts, and references will be noted here. "9:1410," for example, refers to test 1410 in the *Ninth Mental Measurements Yearbook*; "T4:3010" refers to test 3010 in *Tests in Print IV*.

[420]
The Hypothetical Test: Reading.
Purpose: Designed to "measure achievement in reading."
Population: Grades 9–12.
Publication Dates: 1989–1994.
Acronym: HYPE.
Scores, 3: Vocabulary, Comprehension, and Total.
Administration: Individual or group.
Forms, 3: Survey, Abbreviated, Complete Battery.
Restricted Distribution: Distribution of Survey Form restricted to school principals.
Price Data, 1995: $70 per complete kit including 100 tests, scoring key, and manual ('94, 120 pages); $9 per scoring key; $32 per manual.
Special Editions: Braille edition available.
Time: 50 (60) minutes.
Comments: May be self-scored.
Author: Jane J. Doe.
Publisher: Hypothetical Tests, Inc.
Cross References: See T4:3010 (2 references); for reviews by John Roe and Robert Smith of an earlier edition, see 9:1410 (6 references).

Review of the Hypothetical Test: Reading by John J. Smith, Associate Professor of Instruction and Learning, State University, Jonestown, Any State

The actual text of the test review would be here. Space does not permit including a review.

Purpose: A brief, clear statement describing the purpose of the test; often these are quotations from the test manual.

Publication Date: The inclusive range of publication dates.

Acronym: Acronym by which the test may be commonly known.

Scores: The number of explicit scores is presented along with the descriptions of what they are intended to measure.

Forms: All available forms, parts, and levels are listed.

Price Data: Price information is reported for test packages, answer sheets, accessories, and specimen sets.

Time: This is the amount of time to take, and administer, the test. The first number is the actual working time examinees are allowed, and the second (parenthesized) number is the total time needed to administer the test.

Comments: Special notations and comments.

Publisher: The publisher's full address can be found in the Publishers Directory and Index.

Source: Entry information is from "*Mental Measurements Yearbook* and *Tests in Print: A Guide to the Description Entries.*" Lincoln, NE: Buros Institute of Mental Measurement. Used with Permission.

Source: Entry information is from *Mental Measurements Yearbook and Tests in Print* and *Tests in Print: A Guide to the Descriptive Entries.* Lincoln, NE: Buros Institute of Mental Measurement. Used with permission.

LOCATING COMPUTERIZED TESTING MATERIALS

General Internet Searches

You may have some luck searching for computerized testing products using Google or another search engine. Be sure to use "computer testing + education" as the search term; otherwise, you will get lots of false links. Most sites you find will be selling computerized testing products, so you cannot find objective expert evaluations of products at these sites.

Test Publishers' Websites

Test publishers may have several computerized testing products, which will be listed on their

Websites. Again, these sites are marketing tools, so you will not find objective evaluations of the products at the sites. A list of test publishers and their URLs is given in Appendix L of this book.

LOCATING UNPUBLISHED TEST MATERIALS

Not all test materials are published. Many tests have been used in research and evaluation projects. Some of these are available, if you can find them. We describe some sources for locating these types of tests in this section.

Locating Unpublished Tests from Print Sources

ETS Test Collection Earlier we discussed the online *ETS Test Collection* as an online database of tests. Many of the tests in this database are unpublished. Look for the term *unpublished* in the description. The ETS Test Collection has incorporated the former **Tests in Microfiche** collection, which contains more than 800 unpublished tests used in education, business, and psychology. These tests are now downloadable from the ETS Test Collection database.

Directory of Unpublished Experimental Measures This directory edited by Goldman and Mitchell (2002) lists unpublished tests and surveys in a variety of educational and psychological areas. The listings are arranged in 24 categories and include tests' availability, purpose, content, format, and related research. Each volume has a cumulative index that lists all the approximately 5,000 tests across the earlier volumes.

Locating Unpublished Tests Online

Health and Psychosocial Instruments (HAPI) This resource is available online through Ovid Technologies (a vendor of databases) at http://www.ovid.com/. It includes instruments from journal articles in health sciences, nursing, psychology, and social sciences.

College and University Library Subject Guides Some academic libraries in colleges and universities provide online subject guides in testing and measurement. These may include links to bibliographies of unpublished tests and survey instruments. Two examples are: *Tests and Testing Information* (Corby, 2002) at Michigan State University (http://www.lib.msu.edu/corby/psychology/) and *Test and Measures in the Social Sciences: Tests Available in Compilation Volumes* (Hough, 2005) at the University of Texas Arlington (http://libraries.uta.edu/helen/lests&meas/testmaniframe.htm/). You should check your local library for *subject guides on testing and measurement*.

ETS Test Collection We discussed earlier this database of approximately 20,000 tests and other assessment instruments. It contains information on both published and unpublished instruments. You can access the *ETS Test Collection* database directly at http://www.ets.org/testcoll.

General Internet Searches If you are unable to locate a test through one of the preceding sources, you could try searching the Internet with the test title, author's name, or subject area. As we mentioned earlier, including "test" or "assessment" with the subject-area key word helps to narrow your search. Usually, searching through ERIC will yield more relevant hits for educational tests than conventional search engines. You may also try searching on *PsycINFO*.

RESTRICTIONS ON PURCHASING AND USING TESTS

Purchasing Restrictions

Although you may find the name of a test you want to use, its availability may be restricted. To guard against assessment abuse, some publishers restrict the sale of test materials. The publisher's catalog will list any restrictions on test purchasing: The sale of certain tests, especially individually administered intelligence and personality tests, is restricted to qualified psychologists. Typically, publishers label the tests according to the severity of the restrictions in purchasing:

Level A—may be ordered on official letterhead by an agency or organization in which qualified persons will administer and interpret the results. The agency or institution would employ persons who meet the recommendations of the *Standards for Educational and Psychological Testing* (AERA et al., 1999). An individual who is ordering would have to verify completion of sufficient training and a course in test interpretation and use from a recognized program.

Level B—individuals will need to verify that they have had sufficient graduate-level training (typically a master's degree) and supervised experience to administer and interpret the test being ordered. Membership in an appropriate professional association may be required. The recommendations of the *Standards* would be followed.

Level C—individuals need to verify that they have a PhD or related degree in psychology or education as well as appropriate coursework and supervised training in administration and interpretation of the test being ordered.

Sales restrictions vary with the publisher, each implementing a somewhat different policy on establishing a purchaser's qualifications and selling tests. The test user must be sure to acquire the requisite training and experience before purchasing a test. A form needs to be completed, signed, and submitted to the publisher for approval before a test can be purchased.

If you are a practicing teacher and want to review the achievement tests your school district uses, then you should contact the district's testing director. If you are taking a course in which you are expected to write an evaluation of a test, you should ask your instructor how to proceed. Your university may have a testing office that contains specimen sets of tests for this purpose. If you must order a test, your instructor will likely need to prepare a letter for you to include with the order explaining the assignment and how the test will be used. Start your search for a test early. Last-minute searches will likely result in problems completing your assignment.

The previously mentioned **Standards for Educational and Psychological Testing** gives the test publisher the responsibility for telling the user the qualifications needed to interpret test results properly. However, *Standards* (AERA et al., 1999) states, "Those who use psychological tests should confine their testing and related assessment activities to their areas of competence, as demonstrated through education, supervised training, experience, and appropriate credentialing." (Standard 12.1, p. 131).

Guidelines for Proper Test Development and Use

Test Standards A useful publication for evaluating educational and psychological tests is the *Standards for Educational and Psychological Testing* (AERA et al., 1999). Prepared jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, the *Standards* describe various kinds of information that a publisher should provide in a test manual and accompanying materials. It includes suggestions for how a test should be developed as well as guidelines for how a test should be used. Further information can be obtained by calling the National Council on Measurement in Education.

The Code The ideas and concepts in the *Standards* are directed to the professional tester rather than to measurement students and the public. However, the Joint Committee on Testing Practices (2004) prepared a set of major obligations for professionals (like yourself) who use tests in formal educational testing programs. These obligations, described in the *Code of Fair Testing Practices in Education (Revised),* are included in Appendix B. The code will be especially useful to you as you evaluate a test or your school's testing program. The code lists separate obligations for test users and for test developers. The *Code of Professional Responsibilities in Educational Measurement* (NCME, 1995) also describes professional obligations and is reproduced in Appendix C. (See also Chapter 5.)

EVALUATING AND SELECTING A TEST

Because tests play important roles in the educational system, school officials should select them carefully. Before your district adopts a test, a committee of parents, teachers, and administrators should carefully examine and evaluate it. This section describes a systematic procedure for conducting such a review and evaluation. Part of your professional responsibility as a teacher is to participate and offer informed judgments when serving on such test selection committees. School administrators who select tests without the informed judgment of teachers run the risk of egregious errors. Because no test can perfectly match a school district's needs, comparing the merits of one test with another is an important step in choosing the better product.

Clarify Your Purpose

The first step in reviewing a test is to pinpoint the specific purpose(s) for obtaining student information and to find out who will be using the information to make decisions. The more specific you are about the purposes and conditions under which assessment information will be used, the better you will be able to select the appropriate procedure. Information from Chapter 3, which discusses test score validity, will be especially helpful.

Things you need to keep clearly in mind as you begin your selection include:

- *The school setting in which the assessment will be used*—type of community, ages or grades of students, persons who will be helped by an appropriate assessment, and persons who will be in charge of using the assessment results.
- *The specific decisions, purposes, and/or uses intended for the assessment results*—for example, identifying specific reading skills needing remediation, appraising students' emotional needs or areas of anxiety as a prelude to counseling, appraising students' aptitude for mechanical activities that a counselor will discuss during guidance sessions, or surveying general levels of reading and mathematics achievement to report curriculum evaluation information to the school board.
- *The way you believe that using test scores or other assessment information will help improve the decision, serve the purpose, or solve the problem*—the better you can articulate, from the outset, what you expect an assessment procedure to accomplish, the better you will be able to evaluate the many options open to you and to choose the most satisfactory one.
- *The need to strike a balance between the strengths and limitations of performance tasks relative to multiple-choice tests*—such factors as time, cost, in-depth assessment of narrow curricular areas, and less in-depth assessment of broad areas of the curriculum. The assessment procedure you select will be the result of compromises on several dimensions, so it is helpful to think about these early in the process.

Put the New Assessment Plans into Local Context

Before you set out to select a new assessment procedure, you should take stock of the assessments already being used in the district. For example, what type of assessments do teachers already do, of what quality are these assessments, and do they serve the perceived need?

External Assessments Versus Teacher-Made Assessments You will need a perspective on what an external assessment contributes beyond the school-based assessments currently used by teachers. Externally imposed assessments do not match a local curriculum framework exactly. You may decide, for example, that it will be wiser and instructionally more effective to spend the district's money in professional development for improving teacher-made assessment procedures rather than purchasing an **external assessment procedure** such as a standardized test. *In general, a school district should rely on teacher-made assessments for 90% to 95% of its assessment needs*. Principals, because they are responsible for the quality of the instruction in their schools, bear a special responsibility to evaluate teacher-made assessments to ensure they are of high quality and should be aware of how high-quality, teacher-made assessments impact students' learning.

State-Mandated Assessments Versus Standardized Tests States have mandated standardized assessment programs. These programs may be basic skills assessments, accountability programs, or more complex assessments. To reduce redundancy, the assessment a school district purchases should supplement the mandated assessment and serve other, nonduplicating purposes. There has been an increase in mandated state assessments following the federal NCLB legislation. Content and performance standards have been defined, and states are required to attend to these to participate in federal funding. Chapter 15 gave a set of guidelines for selecting a standardized achievement test that is compatible with state-mandated assessment requirements.

Instructional Value of Standardized Tests As discussed in Chapter 15, standardized assessments with norms and educational development scales are most helpful to (a) assess students' relative strengths and weaknesses across curricular areas, (b) assess students' growth within a specific curricular area, and (c) provide an "independent, external" assessment of students' accomplishments relative to a standardization sample. You should weigh these purposes against teacher-based assessments; use instructional benefit to students as a criterion.

Evaluating a School District Sometimes a school district wishes to use an external assessment, such as a standardized test, to evaluate itself. School officials should be aware that not only do single tests provide an especially poor foundation on which to evaluate teachers and curricula, but also that program evaluation itself is a technical area requiring well-prepared professional evaluators. Very often, qualified program evaluation personnel are not on a district's payroll. Being unaware of the need for a professional program evaluator, school officials often assign the task to persons professionally trained in other areas, such as school psychologists or guidance counselors. Superintendents wanting to use assessments for program evaluation may wish to consult curriculum evaluation experts before designing these evaluation strategies. One suggestion is to contact the American Educational Research Association (http://www.aera.net) and ask about contacting a member of Division H who lives near your school district. Another organization you might contact is the American Evaluation Association (http://www.eval.org/).

Figure 17.5 summarizes some factors affecting the difficulty of the school's educational task. Information about these factors should be used along with test results to help interpret a school's effectiveness.

Qualifications of the Staff Another consideration is the qualifications of a school district's staff in relation to the assessment procedure proposed. For example, specially trained professionals are needed to administer and interpret individual intelligence and personality tests, as well as group-administered scholastic ability tests. If such professionals are in short supply in a district, you will want to use other assessment procedures. Similarly, using performance assessments and portfolios requires educating teachers about scoring and interpreting these procedures. This will cost time and money that a district may not have. Sometimes partial implementation may be helpful, such as assessing students at some grades and not others.

Review the Actual Assessment Materials

Obtain Copies of the Test to Review After locating potential assessment instruments and reading reviews if available, narrow your choices to a few assessment procedures that appear to suit your needs. Obtain copies of the assessment materials and tasks; detailed descriptions of the assessment content and rationale behind its selection; materials related to scoring, reporting, and interpreting assessment

FIGURE 17.5 **Facts to be reported in addition to standardized test results when evaluating school effectiveness.**

Attendance Includes absences of staff and students from school and parents from participation in parent-teacher organizations.

Holding power Includes graduation and dropout rates.

Parent involvement Includes parent-teacher organizations, volunteers, and parent-staffed programs.

Diversity Includes staff and student gender, ethnicity, and home language and staff responsibilities.

Economic conditions Includes parent income levels and students receiving free or reduced-cost lunches.

Stability Includes percent of staff and students new to a school district.

Experience Includes years of teaching experience and years of education beyond the initial qualifications.

Staff development Includes in-service programs, peer mentoring, collaboration with businesses or colleges, and courses taken.

Programs for students Includes study skills, counseling, dropout and at-risk prevention, reentry, cross-age tutoring, extracurricular, and summer school.

Achievement Includes performance of students at the next higher educational level; longitudinal patterns of achievement test results, student awards and honors, per student library loans, National Merit scholars, college entrance test results, and out-of-class student accomplishments.

School environment Includes incidents of vandalism and violence, gang-related activities, types of disciplinary actions, special services, extracurricular activities, and library facilities.

Instructional variables Includes length of day, year, and class periods; amount of time per subject per week; number of students using extended day academic program; homework actually assigned; and percent of school days devoted exclusively to academic learning.

Fiscal Includes average teacher, staff, and administrator salaries; expenditures per student.

Source: Adapted from "Putting Test Scores in Perspective: Communicating a Complete Report Card for Your Schools," by K. K. Matter, in *Understanding Achievement Tests: A Guide for School Administrators* (pp. 121–129) by L. M. Rudner, J. C. Conoley, and B. S. Plake (Eds.), 1989. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

results; information about the cost of the assessment materials and scoring service; and technical information about the assessment. (See our earlier discussion about restrictions when ordering tests.)

Much of this material may be bundled together in a **specimen set**, which is designed as a marketing tool as well as for critical review of materials. As a result, not all materials you will need to review a test intelligently are included. For example, some publishers' specimen sets do not include complete copies of the assessment booklets or scoring guidelines. You will need to order these separately.

Technical Information About a Test's Quality

Technical information about a test's quality is not found in a publisher's catalog. A test's **technical manual** gives information about how the test was developed, reliability coefficients, standard errors of measurement, correlational and validity studies, equating methods, item analysis procedures, and norming-sample data. Technical manuals are not typically included in specimen sets and must be ordered separately from the tests. Often the publisher prepares several technical reports for a standardized test. Although school testing directors should have copies of the technical manuals for the tests the school uses, too often they do not. Some colleges and universities that maintain test collections for their faculty and students may have technical manuals. Usually, you will need to order the technical manual directly from the test's publisher.

The Committee Should Review All Materials

Once you obtain the materials, the committee can review them. Be sure to compare similar assessments against the purposes you had in mind for using the assessments. It might be helpful for the committee to obtain input from noncommittee members for certain parts of the assessment: for example, mathematics teachers for the mathematics section, reading teachers for the reading assessment, and so forth. You could also call on a college or university faculty member to help: For example, a testing and measurement faculty member may be better qualified to review and/or explain technical material. Contact the National Council on Measurement in Education (http://www.ncme.org) for the names of specialists who live near your school district.

Achievement Tests Must Match the Curriculum

It is important to match each test item with your state's standards and state or local curriculum. You do this by obtaining the complete list of standards or learning targets, organized by grade level. Two persons independently read each test item and record which standard or learning target it matches. When all items have been matched, the persons compare their results and reconcile the differences. The findings are summarized in a table that lists each standard and the ID number of the test items matching each. The number of nonmatching items is also recorded. This should be done separately for each grade, because a test's items may appear at a grade level that is different than the grade at which the corresponding learning target is taught. If there are a lot of these grade-sequence mismatches, the test will not be suitable for your school district. Be sure to note especially the match between the kinds of thinking and performance activities implied by the standards and the test items. Often the content matches, but the thinking processes and performances required do not. An example of how to do this is found in Nitko et al., (1998). As we discussed in Chapter 3, one should examine whether a test and the standards are aligned with respect to content span, depth of understanding required, topical emphasis, expected student performance, and applicability of the test for all students (La Marca et al., 2000).

Finally, find out the month during which the district plans to administer the test. Then, determine what proportion of the test's items assess content that will have been taught before testing begins. When a test assesses content students have not yet been taught, scores are lowered. (See Chapter 16 for further discussion of this point in connection with grade-equivalent scores.)

Pilot the Test If possible, you should administer the assessment to a few students to get a feel for how students might respond. This would be especially important with writing tasks or performance tasks. You may find that for some otherwise appealing performance tasks, student time limits or instructions are not sufficient and confusion results. This is less likely if the assessment was professionally developed and standardized on a national sample.

A Sample Outline for Your Test Review

It will help your review if you systematically organize relevant information in one or two pages. Using a form is a concise way of sharing information among committee members or with others who may help make decisions about the choice. Figure 17.6 suggests what information to record for your review in such a form. You will find an online lesson in how to use this form along with reviews from the *MMY* to systematically review a test in Nitko (2005a) at http://www.unl.edu/buros/bimm/html/lesson02.html.

FIGURE 17.6 Suggested outline for recording relevant information for reviewing and evaluating an assessment procedure.

Identifying information
1. Title, publisher, copyright date
2. Purpose of the test as described by the publisher
3. Grade level(s), subject(s), administrative time
4. Cost per student, service costs
5. Types of scores and norms provided

Content and curricular evaluation
1. Publisher's description and rationale for specific types of tasks
2. Quality and clarity of the tasks themselves
3. Currency of the content and match to recent curricular trends
4. Match of the tasks to each of the school district's curricula
5. Inclusion of ethnic and gender diversity in the task content

Instructional use evaluation
1. Publisher's description and rationale for how the assessment results may be used by teachers to improve instruction
2. Local teachers' evaluations of how the assessment results could be used for improving their instruction
3. Overlap of assessment with the existing teacher-based assessment procedures

Technical evaluation
1. Representativeness, recency, and local relevance of the national norms
2. Types of reliability coefficients and their values (use average values if necessary)
3. Summary of the evidence regarding the validity of the assessment for the purpose(s) you have in mind for using it
4. Quality of the criterion-referenced information the assessment provides
5. Likelihood that the assessment will have adverse effects on students with disabilities, minority students, and female students

Practical evaluation
1. Quality of the manual and teacher-oriented materials
2. Ease of administration and scoring
3. Cost and usefulness of the scoring services
4. Estimated annual costs (time and money) if the assessment procedure is adopted for the district
5. Likely public reaction to using the assessment procedure

Overall evaluation
1. Comments of reviewers (e.g., *MMY* or *Test Critiques*)
2. Conclusions about the positive aspects of the assessment
3. Conclusions about the negative aspects of the assessment
4. Summary and recommendation about adoption

List of references and sources used

HOW A STANDARDIZED TEST IS DEVELOPED

To be an informed consumer of assessments, you need to be aware of the steps a publisher should follow when developing a test, as well as the activities involved in each of these steps. Your judgment of how well the publisher carried out each step should be part of your evaluation of the quality of an assessment procedure.

A standardized test should be the product of a carefully conducted program of research and development. The activities involved in each step are briefly described in this section. Such a well-run development program involves the work of many persons and includes the following steps (Robertson, 1990, pp. 62–63).

1. Assemble preliminary ideas.
2. Evaluate proposal (approve/reject).
3. Make formal arrangement (sign contract if publication is approved).
4. Prepare test specifications.
5. Write items.
6. Conduct item tryout.
   a. Prepare tryout sample specifications.
   b. Prepare participants.
   c. Prepare tryout materials.
   d. Administer tryout items.
   e. Analyze tryout data.
7. Assemble final test form(s).
8. Conduct national standardization.
   a. Prepare standardization sample specifications.
   b. Obtain participants.
   c. Prepare standardization materials.
   d. Administer tests.
   e. Analyze data.
   f. Develop norms tables.
9. Prepare final materials.
   a. Establish publication schedule.
   b. Write manual.
   c. Prepare test books and answer forms.
   d. Manufacture/produce/print materials.
10. Prepare marketing plan.
    a. Initiate direct mail promotion.
    b. Initiate space advertising.
    c. Train sales staff.
    d. Attend professional meetings and conventions.
11. Publish.

More details about these steps and what test developers are expected to do at each stage of development are found in the *Standards for Educational and Psychological Testing* (AERA et al., 1999).

Note, however, that many assessments available in the marketplace do not follow all steps, because to do so is quite expensive and time-consuming. The steps most likely to be omitted are those concerned with collecting and analyzing data used to improve the quality of the test and/or to support the validity of the claims made for the test. Shortcutting assessment development steps usually means lowering validity, so beware of poorly developed assessments.

Universal Design Considerations

More and more, publishers of large-scale assessments try to incorporate principles of universal design into their test development. **Universal design** is a concept that began in the field of architecture and is intended to maximize access. So for example, when the curb is cut to street level at intersections, people in wheelchairs don't need special assistance to navigate them. However, the curb design is also good for others: people with strollers, rolling briefcases, even the temporarily tired pedestrian. Universal design has rapidly spread to other fields, including educational assessment. The idea is to design tests that work for most test takers, and avoid as much as possible the need for special accommodations.

Applied to educational testing, universal design in assessment means developing tests from the beginning in order to allow the broadest possible range of students to participate. The National Center on Educational Outcomes (NCEO; Thompson, Johnstone, & Thurlow, 2002) lists the following elements of universally designed assessments:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Readers who would like further details about these elements should consult the NCEO report, available at http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

CONCLUSION

In this chapter, we presented a brief overview of resources for locating, evaluating, and selecting published and unpublished tests. Then we described, again briefly, the general process of test development. These overviews should be sufficient to get you started when the occasion arises that you need to find and evaluate an assessment.

In the next chapter, we conclude our consideration of standardized tests with a tour of some of the kinds of tests, in addition to achievement tests, that you may find used in your school. These include scholastic aptitude, career interest, attitude, and personality tests.

EXERCISES

1. Describe the types of assessment information you would find in each of the following sources:
   a. *Mental Measurements Yearbook*
   b. *Tests in Print*
   c. *Tests*
   d. *Test Critiques*
   e. *ETS Test Collection*
   f. *Standards for Educational and Psychological Testing*
   g. a test publisher's catalog
   h. http://www.ncme.org
2. Describe how each of the following professional organizations may help you obtain test information:
   a. Association for Assessment in Counseling and Education
   b. International Reading Association
   c. National Association of Test Directors
   d. National Council on Measurement in Education
3. Read each of these statements and identify the one source that would most likely contain the information the speaker is requesting.
   a. "I want to know what kinds of instruments are available to assess attitudes of female students toward work, home, marriage, and family life."
   b. "I want to know what professionals in the field think of this criterion-referenced test."
   c. "What services does a publisher provide for interpreting assessment results, and what are the charges?"

d. "I want to know the newest instruments developed for assessing perceptual-motor development of primary school students."

4. Suppose you have already located a particular test and you want the specific information about it implied by the following statements. What source would you consult first for each statement?
   a. "What are the reliability coefficients for their test?"
   b. "What kind of norms does the publisher provide?"
   c. "How do test specialists view the quality of the procedures the publisher followed when developing the test?"
   d. "What research studies and reports have used this assessment instrument?"
5. Using the procedures described in this chapter, locate a specific standardized assessment instrument you believe can serve a purpose you have identified.
   a. Then, following the procedures described in this chapter, review and evaluate this assessment instrument in relation to your stated purpose. Write your review and evaluation using the outline given in Figure 17.6, using headings and subheadings appropriately.
   b. Share your evaluation with others in this course.

Scholastic Aptitude, Career Interests, Attitudes, and Personality Tests

KEY CONCEPTS

1. Scholastic aptitude assessments describe a learner's general intellectual skills, rather than specific school achievements. However, intellectual skills are also learned, largely at school, and do not represent innate intellectual "capacity."
2. There are several different types of group-administered scholastic aptitude tests.
3. Other types of group-administered specific aptitude tests include readiness tests, high school and college admissions tests, and tests of aptitude for specific subjects.
4. Individually administered tests of general school aptitude include the *Stanford-Binet Intelligence Scale*, the *Wechsler Intelligence Scales,* and the *Kaufman Assessment Battery for Children*.
5. Assessment of adaptive behavior focuses on how independently students can care for themselves and cope with the demands of everyday life.
6. Vocational and career interests are preferences for specific kinds of activities, and can be assessed with a questionnaire.
7. Attitudes can also be assessed with a questionnaire.
8. A variety of techniques have been developed to measure various aspects of personality.

IMPORTANT TERMS

abstract/visual reasoning subtests
adaptive behavior
affective saliency
age-based scores versus grade-based scores
aptitude (versus achievement)
attitudes
completion test of personality
direction and intensity of attitude
empirically keyed scales
expressed interests, inventoried interests, manifested interests, tested interests
figural reasoning
forced-choice item format
general versus specific intellectual skills
interests
mental age
multiple-aptitude tests
nonverbal tests
omnibus test
people-similarity rationale versus activity-similarity rationale
pictorial reasoning
projective hypothesis
projective personality test techniques
quantitative reasoning
readiness test

From Chapter 18 of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

short-term memory subtests

spiral format

standard age score (*SAS*)

structured (self-report) personality assessment techniques

two-score test

values

verbal comprehension tests

verbal reasoning tests

vocational interest inventories

APTITUDES FOR LEARNING

General Versus Specific Intellectual Skills

Assessments of the kind discussed in this chapter describe a learner's **general intellectual skills**, rather than describing the **specific intellectual skills** a learner needs in, say, next week's geometry lessons. When the knowledge or skills students need for an upcoming lesson are specific and narrow, the students' present level of knowledge and skills are the best predictor of their learning success. For most of these specific, day-to-day instructional decisions, you will have to develop your own assessment procedures.

General Intellectual Skills and Aptitudes Measurement

A student's past performance in a specific course is not very helpful in establishing expectations for learning new material whenever (a) the student must learn to perform in ways that are quite different from those learned in the past, (b) the student's past performance has been very erratic, (c) previous test scores or school grades are known to be very unreliable or invalid, or (d) the student's record of past performance is not available.

Consider a ninth grader, for example, who wants to study Spanish for the first time, having had no previous foreign language training or experience. A test of Spanish language knowledge provides no information about this student's chances of succeeding in an upcoming Spanish course. In such cases, an assessment of more general intellectual skills and abilities *related to language learning* will better predict success. Usually such tests assess English language skills and concepts, acquired auditory learning skills, and applied memory skills. Similarly, a student transferring from another school system or moving from one educational level to the next may have complete records, but the meaning of these records may be unclear. To clarify them, a school may test the student with an instrument assessing broad intellectual skills or scholastic aptitude.

School officials can use a number of ways to predict a student's likely success in an educational program. Three examples are the student's (a) level of past achievement for the same specific type of performance as the new performance the student needs to learn, (b) level of general scholastic ability, and (c) ability in several specific aptitudes related to the new performance to be learned. The validity of these predictors is related to the specificity of the performance the school wants to predict. If a school wants to predict a very specific performance (for example, solving quadratic equations), then (a) a student's prior achievement of a very similar kind is the best predictor, (b) a student's general scholastic aptitude is the next best predictor, and (c) assessment of specific aptitudes is least preferred. If the school wants to predict a very general performance (such as overall school performance as measured by first-year-student grade point average), then the preferred order of predictors is (a) general scholastic aptitude assessment, (b) assessments of prior specific achievement, and (c) assessments of specific abilities (Snow, 1980).

Aptitude Tests Measure Learned Behavior

Capacity Is Not Fixed Tests assessing aptitude or intellectual skills reflect only past learning. They do not directly assess innate ability or "capacity." Further, because we cannot obtain a sample of performance from the future, they cannot directly assess future ability. We have to be content to use past and present learning to predict future learning. It is important to recognize that a student's "aptitude for learning" implies learning through a specific type of instructional approach. If you change the instructional approach drastically, the student's aptitude for learning changes as well. A student's aptitude is influenced also by a number of facts of development (including biological makeup), experience in the environment (including interactions with other persons), and a complex interaction of the two.

The very idea of "capacity" places some upper limit on a student's ability to learn. This limitation is likely to be untrue in general. For example, a student's capacity to do algebra may depend on the way a teacher currently teaches it, on the mathematics concepts the student learned previously, and on the motivational level a teacher stimulates in the student, as well as on some kind of native endowment. It seems reasonable to conclude that both developmental (life history) and instructional conditions affect one's particular potentials.

Aptitude-Achievement Distinctions It sometimes troubles teachers to see assessment instruments bearing titles such as *readiness, intelligence, general mental ability,* and "*aptitude for* X," but containing items closely resembling those found on achievement tests. It is important to distinguish between the abstract concepts of **aptitude** and **achievement** and the observations we make to infer the state of a person's aptitude and achievement. We can define an aptitude for *X* as the present state of a person that indicates the person's expected future performance in *X* if the conditions of the past and present continue into the future (see Carroll, 1974). A student's present aptitude (or state) could be indicated in many ways.

> Thus: [An] "aptitude test" is only one indicant of aptitude. Other indicants of aptitude could include scores on achievement tests, data on prior performance in activities similar to those for which we wish to predict success, and information derived from procedures for assessing personality, interest, attitude, physical prowess, physiological state, etc. (Carroll, 1974, p. 287)

Scholastic aptitude tests deliberately set out to assess a student's reasoning rather than the student's recall of factual knowledge or ability to use well-learned rules on problems practiced in school. These tests differ from traditional standardized achievement tests in at least three ways:

> First, tests of reasoning ability, especially mathematical reasoning, require a relatively small declarative knowledge base. The . . . amount of mathematical knowledge required by the typical SAT [for example] is rarely beyond that taught in a first year high school algebra course and an introductory semester of geometry. . . . [The SAT places heavy demands, however, on] procedural knowledge or, more precisely, the procedural use of declarative knowledge. . . .
>
> A second way in which reasoning tests differ from subject matter tests is in the quite deliberate way in which they were constructed to not depend upon specific subject matter content. The verbal reasoning skills measured by the SAT-V, for example, have no specific secondary school course sequence on which they can be referred. A final way in which verbal and mathematical reason tests differ from at least some achievement tests is in degree of problem-solving and reasoning, as distinct from simple memory. Tests in subject matters such as geography, foreign languages, and history make primary demands on memory but minimal demands on problem-solving skills and reasoning. (Bond, 1989, pp. 429–430)

Teaching Conditions for Aptitude Development

Nonadaptive Teaching An important part of the definition of aptitude given earlier was the continuance of past learning conditions into the future. One thing that makes aptitude tests useful predictors of future school success is that schools are generally not very adaptive to individual learners. Thus, the conditions under which students learn this year are usually quite similar to last year's learning conditions.

Underemphasizing Adaptive Teaching When a student must learn under the same conditions from year to year, there is a danger that teachers will believe that the results of the student's scholastic aptitude testing determine what the student is able to accomplish. That is, once they see a student's present aptitude level, they will do little to modify learning conditions to improve the student's aptitude. Past psychological conceptions of the learner have led some educators to overemphasize the (a) consistency of the general scholastic ability of learners, (b) passivity of learners as receivers of information, and (c) categorical placement of learners into educational tracks with narrow ranges of instructional options. They have *underemphasized* the (a) adaptivity and plasticity of learners, (b) learners' ability to actively construct information during problem solving, and (c) responsibility of educational systems to adapt to learners' initial performance levels (Glaser, 1977).

Invalidity for Instructional Placement Unfortunately, the aptitude tests described in this chapter have not been validated for use in assigning students to different kinds of instructional methods. Rather, they have been built to predict how well students will perform when they must adapt to the fixed type of instruction. You should view the tests' helpfulness for decision making in that light.

Scholastic Aptitude Test Score Stability

Importance of Score Stability An important school concern is whether a student's scholastic aptitude remains constant or stable over time. If a student's scholastic aptitude test score changed erratically every year, you would have no confidence that it assessed a useful characteristic. Although on a student-by-student basis scores may systematically rise or decline, a definite tendency exists for students to maintain similar, but not identical, ranking in their age group throughout their school years. In general, changes in students' rankings on general scholastic aptitude tests tend to be greater (a) as the time interval between the two testings grows and (b) the younger the students were at the time of the initial testing. Although groups of students tend to maintain their relative position in the distribution of aptitude scores, important changes in individual students do occur. Therefore, if a school wants to use a student's scholastic aptitude score for guidance or placement decisions, it should bear in mind that there are sufficient differences in individual students' patterns of score change to justify reassessment each time a decision is made (Sattler, 1988).

Factors for Score Stability Among the factors that work to keep students' rankings on aptitude tests about the same over time are (Anastasi, 1988; Sattler, 1988):

1. The genetic makeup of students remains stable.
2. If a student's socioeconomic level, family configuration, and sociocultural influences remain stable over a long period, these contribute to aptitude stability.
3. Development and prerequisite learning is rarely reversible, so earlier development and learning continues to exert similar impact on new development and learning.
4. If the content assessment by different scholastic aptitude tests is similar, students' scores will be similar from one testing to the next.

Reasons for Score Changes Among factors that work to change students' rankings on aptitude tests from one testing to another are the following:

1. *Errors of measurement*—Even if the person's "true score" were to stay the same, the obtained score is likely to be different due to a test's unreliability (see Chapter 4).
2. *Test differences*—The content of tests produced by different publishers will vary. Also, the content of the same publisher's test may vary with the age level of the student taking the test. Tests designed for young children are more concrete and perceptual; those designed for older children are more abstract and verbal.
3. *Norm-group differences*—The norms of different publishers' tests are not comparable. Because mental ability scores are norm-referenced, differences in scores may be due to differences in norms.
4. *Special interventions and enriched environments*—If a person's environment dramatically and persistently becomes more intellectually nurturing, that person's scores on a scholastic aptitude test are likely to increase. Conversely, if the person becomes physically or emotionally ill or deprived in a way that interferes with intellectual development, then aptitude scores may decrease.

GROUP TESTS OF SCHOLASTIC APTITUDES

Types of Group Aptitude Tests

Advantage The principal advantage of group testing over individual testing is the efficiency and cost savings gained by testing many persons at the same time. The ease with which group tests can be administered and scored has contributed greatly to schools adopting them.

Number of Aptitudes Reported There are different types of group aptitude tests. The **omnibus test** contains items assessing different abilities that comprise general scholastic aptitude, but it provides only a single score. A **two-score test** also assesses several different kinds of specific abilities, but reports only two scores, usually verbal/quantitative or verbal/nonverbal. The items on the verbal section of the test, for example, may assess several kinds of specific verbal abilities, but only one verbal ability score is reported.

Some school ability tests report three scores, such as verbal, quantitative, and nonverbal. **Nonverbal tests** assess how well students process symbols and content that have no specific verbal labels, such as discerning spatial patterns and relations or classifying patterns and figures. **Multiple-aptitude tests** assess several different abilities separately and provide an ability score for each. Multiple-aptitude tests, for example, may provide separate scores for verbal reasoning, verbal comprehension, numerical reasoning, and **figural reasoning**.

Type of Test to Use The type of group scholastic aptitude test a school should use depends on how the staff will use the scores. Multiple-aptitude tests are most useful for providing information that profiles students' strengths and weaknesses to make better decisions about further schooling or planning a career. Omnibus tests are most useful when a school wants an estimate of their students' general level of school ability for purposes of predicting future success under standard classroom conditions. Examples of two-score and multiple-score aptitude tests are given in the following sections. Others are listed in Appendix K.

Two-Score Test: *Otis-Lennon School Ability Test*

Test Content The *Otis-Lennon School Ability Test* (*OLSAT*) provides verbal and nonverbal part scores as well as a total score. The identification of a test item as verbal versus nonverbal depends on whether students must understand English to answer the item. For example, Numeric Inference is classified as quantitative reasoning because English language is not necessary to succeed on the items. Once the directions for the subtest are understood, an examinee could answer the questions without knowing English. Arithmetic Reasoning, on the other hand, is classified as verbal reasoning because it "is made up of verbal problems, does not depend on computation, and depends on understanding English" (Pearson, 2003). Figure 18.1 describes the clusters and types of items the *OLSAT* contains.

FIGURE 18.1 **Description of the kinds of items on the *OLSAT*.**

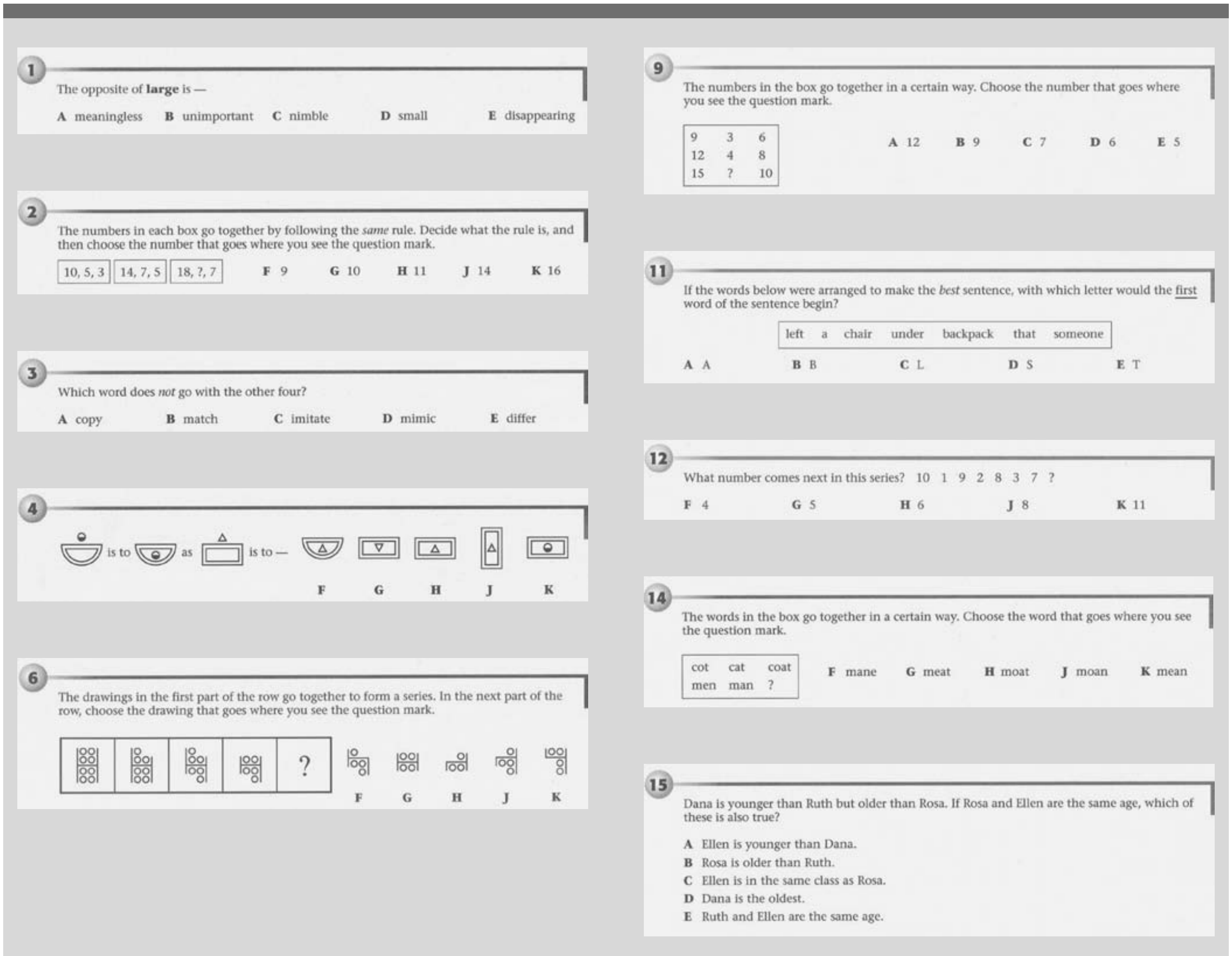
	Cluster description	Types of items
Verbal clusters	***Verbal comprehension*** depends on the ability to perceive the relational aspects of words and word combinations, to derive meaning from types of words, to understand subtle differences among similar words and phrases, and to manipulate words to produce meaning.	Following Directions Antonyms Sentence Completion Sentence Arrangement
	Verbal reasoning depends on the ability to infer relationships among words, to apply inferences to new situations, to evaluate conditions in order to determine necessary versus optional, and to perceive similarities and differences.	Aural Reasoning Arithmetic Reasoning Logical Selection Word/Letter Matrix Verbal Analogies Verbal Classification Inference
	Pictorial reasoning assesses the ability in young children to reason using pictorial representations. These items assess the ability to infer relationships among objects, to evaluate objects for similarities and differences, and to determine progressions and predict the next step in those progressions.	Picture Classification Picture Analogies Picture Series
Nonverbal clusters	***Figural reasoning*** items assess the ability to use geometric figures to infer relationships, to perceive progressions and predict the next step in those progressions, to generalize from one set of figures to another and from dissimilar sets of figures, and to manipulate spatially.	Figural Classification Figural Analogies Pattern Matrix Figural Series
	Quantitative reasoning items assess the ability to use numbers to infer relationships, derive computational rules, and predict outcomes according to computational rules.	Number Series Numerical Inference Number Matrix

Source: Adapted from *Otis-Lennon School Ability Test: Eighth Edition, Assessing the Abilities That Relate to Success in School.* Copyright © 2009 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

Test Organization The *OLSAT* is organized into seven levels: Level A (kindergarten), Level B (Grade 1), Level C (Grade 2), Level D (Grade 3), Level E (Grades 4 and 5), Level F (Grades 6, 7, and 8), and Level G (Grades 9, 10, 11, and 12). Not all of the different types of items are given at every grade. In Grades K through 2, the test items are organized into three sections, each section containing distinct item types. The three-part format places pictorial items together and items with dictated stems together, separated from other teacher-paced but not dictated items. The Grade K through 2 tests also group types of tasks (e.g., classifying) together. In Grades 4 through 12, similar types of items are not grouped together into subtests, but are arranged into a **spiral format**, similar to that shown in Figure 18.2. One item of each type is presented; then the sequence is repeated, but with more difficult items. Items at the upper levels are entirely self-administered: Students read the directions and answer the items without teacher pacing. The Grade 3 test has two sections, a classification section with figural and verbal items spiraled according to the easy-hard format, and then a section with all the rest of the items arranged in the same spiral format as that for Grades 4 through 12.

Norm-Referencing Scheme The publisher of the *OLSAT* uses several norm-referencing schemes to report the verbal, nonverbal, and total test results. These are illustrated on the individual student report shown in Figure 18.3. These types of scores

FIGURE 18.2 **Examples of the type of items on the *Otis-Lennon School Ability Test* (8th ed.).**



1 The opposite of **large** is —

A meaningless **B** unimportant **C** nimble **D** small **E** disappearing

2 The numbers in each box go together by following the *same* rule. Decide what the rule is, and then choose the number that goes where you see the question mark.

10, 5, 3 | 14, 7, 5 | 18, ?, 7 **F** 9 **G** 10 **H** 11 **J** 14 **K** 16

3 Which word does *not* go with the other four?

A copy **B** match **C** imitate **D** mimic **E** differ

4 is to as is to —

F **G** **H** **J** **K**

6 The drawings in the first part of the row go together to form a series. In the next part of the row, choose the drawing that goes where you see the question mark.

?

F **G** **H** **J** **K**

9 The numbers in the box go together in a certain way. Choose the number that goes where you see the question mark.

9	3	6
12	4	8
15	?	10

A 12 **B** 9 **C** 7 **D** 6 **E** 5

11 If the words below were arranged to make the *best* sentence, with which letter would the first word of the sentence begin?

left a chair under backpack that someone

A A **B** B **C** L **D** S **E** T

12 What number comes next in this series? 10 1 9 2 8 3 7 ?

F 4 **G** 5 **H** 6 **J** 8 **K** 11

14 The words in the box go together in a certain way. Choose the word that goes where you see the question mark.

cot	cat	coat
men	man	?

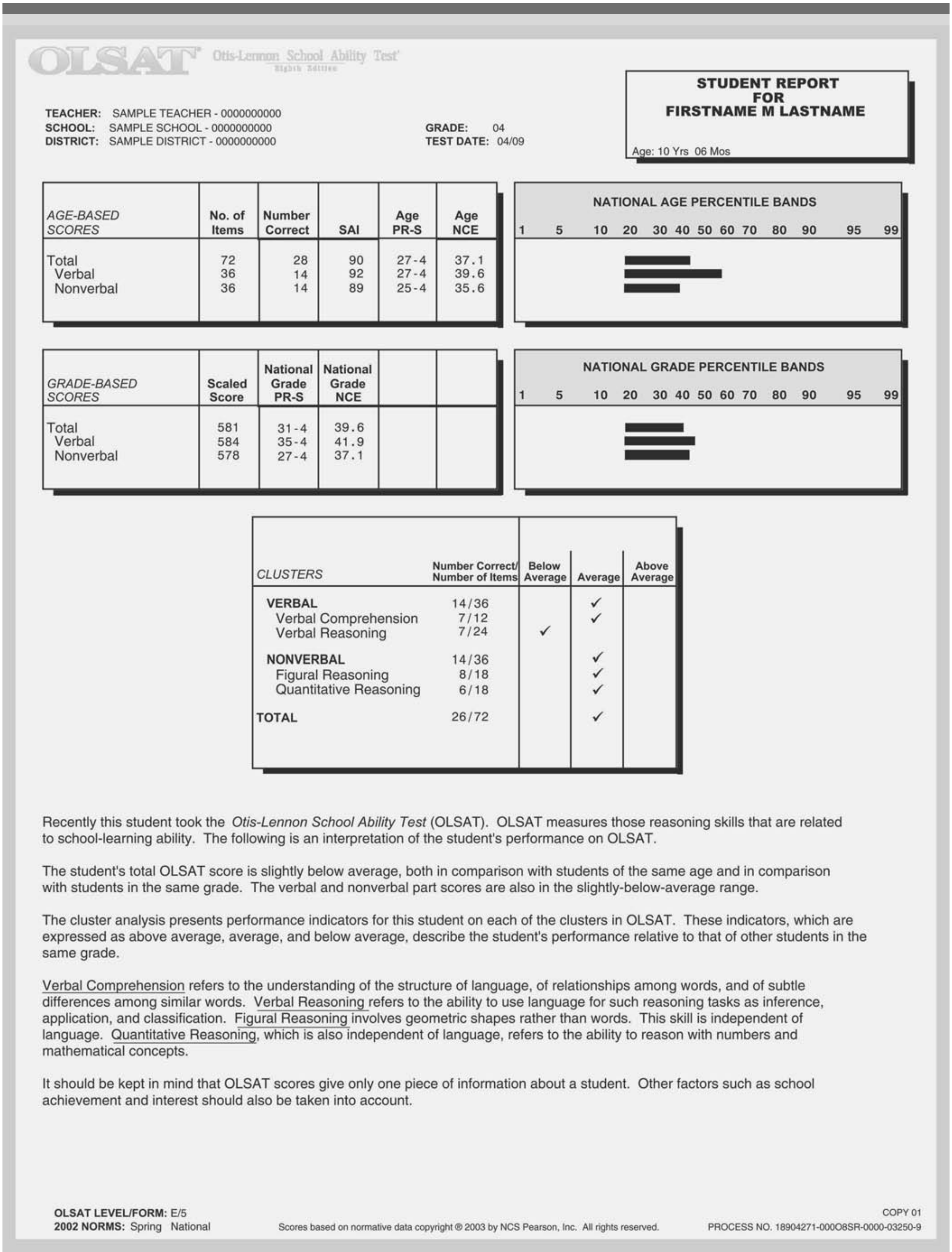
F mane **G** meat **H** moat **J** moan **K** mean

15 Dana is younger than Ruth but older than Rosa. If Rosa and Ellen are the same age, which of these is also true?

A Ellen is younger than Dana.
B Rosa is older than Ruth.
C Ellen is in the same class as Rosa.
D Dana is the oldest.
E Ruth and Ellen are the same age.

Source: Otis-Lennon School Ability Test: Eighth Edition, Level F. Copyright © 2003 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

FIGURE 18.3 **An individual student's report showing the type of scores reported for the *Otis-Lennon School Ability Test*.**



OLSAT® Otis-Lennon School Ability Test®
Eighth Edition

STUDENT REPORT
FOR
FIRSTNAME M LASTNAME

Age: 10 Yrs 06 Mos

TEACHER: SAMPLE TEACHER - 0000000000
SCHOOL: SAMPLE SCHOOL - 0000000000
DISTRICT: SAMPLE DISTRICT - 0000000000

GRADE: 04
TEST DATE: 04/09

AGE-BASED SCORES	No. of Items	Number Correct	SAI	Age PR-S	Age NCE
Total	72	28	90	27-4	37.1
Verbal	36	14	92	27-4	39.6
Nonverbal	36	14	89	25-4	35.6

NATIONAL AGE PERCENTILE BANDS

1 5 10 20 30 40 50 60 70 80 90 95 99

GRADE-BASED SCORES	Scaled Score	National Grade PR-S	National Grade NCE		
Total	581	31-4	39.6		
Verbal	584	35-4	41.9		
Nonverbal	578	27-4	37.1		

NATIONAL GRADE PERCENTILE BANDS

1 5 10 20 30 40 50 60 70 80 90 95 99

CLUSTERS	Number Correct/ Number of Items	Below Average	Average	Above Average
VERBAL	14/36		✓	
Verbal Comprehension	7/12		✓	
Verbal Reasoning	7/24	✓		
NONVERBAL	14/36		✓	
Figural Reasoning	8/18		✓	
Quantitative Reasoning	6/18		✓	
TOTAL	26/72		✓	

Recently this student took the *Otis-Lennon School Ability Test* (OLSAT). OLSAT measures those reasoning skills that are related to school-learning ability. The following is an interpretation of the student's performance on OLSAT.

The student's total OLSAT score is slightly below average, both in comparison with students of the same age and in comparison with students in the same grade. The verbal and nonverbal part scores are also in the slightly-below-average range.

The cluster analysis presents performance indicators for this student on each of the clusters in OLSAT. These indicators, which are expressed as above average, average, and below average, describe the student's performance relative to that of other students in the same grade.

Verbal Comprehension refers to the understanding of the structure of language, of relationships among words, and of subtle differences among similar words. Verbal Reasoning refers to the ability to use language for such reasoning tasks as inference, application, and classification. Figural Reasoning involves geometric shapes rather than words. This skill is independent of language. Quantitative Reasoning, which is also independent of language, refers to the ability to reason with numbers and mathematical concepts.

It should be kept in mind that OLSAT scores give only one piece of information about a student. Other factors such as school achievement and interest should also be taken into account.

OLSAT LEVEL/FORM: E/5
2002 NORMS: Spring National

Scores based on normative data copyright ® 2003 by NCS Pearson, Inc. All rights reserved.

COPY 01
PROCESS NO. 18904271-000O8SR-0000-03250-9

Note: Scores are simulated results.

Source: From *Otis-Lennon School Ability Test: 8th Edition, Score Report Example.* Copyright © 2003 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

are described in the following list. The letters in this description correspond to the letters in Figure 18.3.

1. **Age-based scores** compare this student with norm-group students who are the same age, regardless of grade placement. In addition to raw scores, *OLSAT* reports (1) *School Ability Index (SAI)*, a normalized standard score with mean 100 and standard deviation of 16; (2) *national percentile rank (PR);* (3) *national stanine (S);* and (4) *normal curve equivalent (NCE)*.
2. **Grade-based scores** compare this student with norm-group students who have the same grade placement, regardless of their age. The scores reported in this section are (1) *scaled score,* an expanded scaled score that allows you to track growth in scholastic aptitude over several years because the scale spans all grades; (2) *national percentile rank (PR);* (3) *national stanine (S);* (4) *local percentile rank (PR);* (5) *local stanine (S);* and (5) *local normal curve equivalent (NCE).*
3. *Percentile bands* show the uncertainty interval for the student's scores that are reported in Sections A and B of Figure 18.3. Uncertainty bands are formed by adding and subtracting one *SEM* to the student's score. (See the discussion of *SEM* in Chapter 4.)
4. *Cluster scores* are the raw scores for each of the five clusters at a particular grade level (see Figure 18.1). Below average, average, and above average describe cluster performance in terms of stanines: below average includes stanines 1, 2, 3; average scores fall into stanines 4, 5, 6; and above-average stanines are 7, 8, 9. Stanines are different for the spring and fall standardization groups.
5. *Computer-generated narrative* explains the results in simplified language.

Interpretation of Results Although the *OLSAT* is a two-score test, its authors encourage you to use the total test results as the main interpretive piece of information. They believe that because verbal and nonverbal abilities are needed to succeed in school, the total score is the best overall indicator. They recognize, however, that much of what you teach students relates to verbal learning. Thus if a student is very much higher in nonverbal than in verbal ability, you might be alerted that the student may have good scholastic ability but may have difficulty in highly verbal subjects. Students with higher verbal than nonverbal ability may experience more difficulty with quantitative subjects. The authors recommend that you consider score differences larger than two stanines as meaningful. You should interpret smaller differences much more cautiously because they may represent only measurement error.

Other possible causes for a verbal-nonverbal difference include bilingualism, reading problems, learning disability, hearing impairment, visual impairment, anxiety, illness, or irregularities in test administration. You should request a readministration of the *OLSAT* if a student has a large verbal-nonverbal difference. If retesting verifies a difference and you want further diagnostic information, then you should request assessment with an individual test such as the *Wechsler Intelligence Scale for Children* (described later in this chapter).

Achievement/Ability Comparisons If you administer the *OLSAT* along with the eighth edition of the *Stanford Achievement Test Series* or the *Metropolitan Achievement Tests,* a score called the Achievement/Ability Comparison (AAC) is part of a student's test report. The AAC describes, for each achievement survey battery subtest, how this student's achievement compares to norm-group students who have the same *OLSAT* total score.

To do this, students in the grade-based *OLSAT* norm group are first sorted into stanines. Second, the students within each *OLSAT* stanine are then sorted into stanines for the achievement test subtest (e.g., reading comprehension stanines for all students whose *OLSAT* stanine is 5). Third, within each achievement subtest group from Step two, students are clustered into high (stanines 7, 8, 9), middle (stanines 4, 5, 6), and low (stanines 1, 2, 3) groups. The student's *OLSAT* stanine is reported along with his achievement subtest stanines.

For example, Don's *OLSAT* stanine is 5. Don also took the *Stanford Achievement Test* and scored stanines of 4, 5, and 6 in Total Reading, Total Mathematics, and Spelling, respectively. Compared to the entire norm group for the achievement test, these stanines fall in the middle of the distribution. However, if we look only at those students who attained an *OLSAT* stanine of 5, Don will be in the low AAC range in Total Reading, in the middle AAC range in Total Mathematics, and in the high AAC range in Spelling. These AAC results tell us that compared to other students with the same scholastic aptitude as Don, he is below average in Total Reading, average in Total Mathematics, and above average in Spelling.

Multiple-Aptitude Test: *Differential Aptitude Tests*

Purpose The battery of *Differential Aptitude Tests (DAT)* was originally developed in 1947 to satisfy the needs of guidance counselors and consulting psychologists working in schools, social agencies, and industry (Bennett, Seashore, & Wesman, 1974). The tests were revised in 1962, 1972, 1982, and 1990.

The primary purpose of the tests is to provide information about a student's profiles with respect to different cognitive abilities. This information is used for guiding and counseling students in junior and senior high schools (Grades 7 through 12) as they prepare for career decisions. There are two levels: Level 1 (Grades 7–9) and Level 2 (Grades 10–12). The *DAT* are also used with adults for vocational and educational counseling and as part of a battery of tests for job selection.

Test Content The *Differential Aptitude Tests* report scores for each of the eight subtests shown in Figure 18.4. An additional ninth score, Scholastic Aptitude, which is a combination of the Verbal Reasoning and Numerical Reasoning scores, is reported: This score is used to assess general scholastic aptitude. Figure 18.4 also shows examples of items from each subtest.

Gender-Specific Norms The *DAT* have separate male and female norms, as well as combined norms. Separate norms allow comparisons of a student with his or her own gender, as well as with members of the opposite gender. Cross-gender comparisons may help students consider occupations or educational programs that they would have overlooked. This may surprise you and may seem like a form of gender discrimination. However, because the tests are used for guidance and counseling, this purpose is better served by these separate norms, given the realities of the current job market.

Advantage of the *DAT* An advantage of using a multiple-aptitude battery such as the *DAT* instead of an omnibus or two-score aptitude test is the opportunity it provides for finding some aptitude for which a student has a relative strength. For example, a student may have low general scholastic ability (Verbal Reasoning and Numerical Reasoning) but have high Perceptual Speed and Accuracy or high Mechanical Reasoning. This provides counselors with information on aptitude that they can use to encourage students.

Combining Aptitude With Interest Assessment The *DAT* comes with an optional *Career Interest Inventory*. Using the results of this instrument along with aptitude scores, achievement scores, and school grades can help a student make realistic career or further education decisions. The interest inventory presents sentences describing activities in various types of work and school situations. Students indicate their degree of agreement with the sentences. (Interest inventories are described in greater detail in this chapter.)

GROUP TESTS OF SPECIFIC APTITUDES

The kinds of general scholastic aptitude tests illustrated earlier are widely used in schools, but other types used for special decisions should be mentioned, too. Among these are readiness tests, high school and college admissions tests, and tests of aptitude for specific subjects.

Readiness Testing

Schools often use **readiness tests** as supplemental information to make instructional decisions for first-grade pupils. Often such tests are used to supplement a kindergarten teacher's judgment about a youngster's general developmental and readiness level for first-grade work, especially reading, where grouping by readiness level is a common practice. Because readiness tests measure a child's acquired learning skills, they are frequently classified as achievement tests rather than aptitude tests.

Teachers frequently use readiness tests to help form instructional groups (for example, for reading instruction). When used in this way, they should be considered placement tests. Because teachers use readiness tests to predict implicitly a pupil's likely success in instruction, we discuss them as aptitude tests in this book. We also mentioned them in Chapter 15 as examples of a type of early childhood achievement test that has been experiencing expanded growth in the current NCLB accountability climate.

You should keep in mind the test author's point of view when selecting a readiness test. The author's viewpoint of what constitutes "readiness to learn" will determine the test content (as does an author's viewpoint for every test, of course). If you want to use the scores on a readiness test to make a statement about whether a student has mastered specific prerequisites, you must carefully examine the actual test items to see if they measure the kinds

FIGURE 18.4 **Brief descriptions, time limits, number of items, and a sample item from each subtest of the *Differential Aptitude Tests* (5th ed.).**

Verbal Reasoning (25 min., 40 items)

Measures the ability to see relationships among words; may be useful in predicting success in business, law, education, journalism, and the sciences.

SAMPLE ITEM

Which answer contains the missing words to complete this sentence?

. . . . . is to fin as bird is to . . . . .

A water — — feather
B shark — — nest
*C fish — — wing
D flipper — — fly
E fish — — sky

Numerical Reasoning (30 min., 40 items)

Measures the ability to perform mathematical reasoning tasks; important in jobs such as bookkeeping, lab work, carpentry, and toolmaking.

SAMPLE ITEM

What number should replace R in this correct addition example?

$$\begin{array}{r} 7R \\ +\ R \\ \hline 88 \end{array}$$

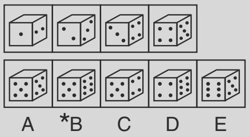
*A 9
B 6
C 4
D 3
E None of these

Abstract Reasoning (20 min., 40 items)

A nonverbal measure of the ability to reason using geometric shapes or designs; important in fields such as computer programming, drafting, and vehicle repair.

SAMPLE ITEM

Choose the Answer Figure that should be the next figure (or fifth one) in the series.



Perceptual Speed and Accuracy (6 min., 200 items)

Measures the ability to compare and mark written lists quickly and accurately; helps predict success in performing routine clerical tasks.

SAMPLE ITEM

Look at the underlined combination of letters or numbers and find the same one on the answer sheet. Then fill in the circle under it.

1 XY Xy XX YX Yy
2 6g 6G G6 Gg g6
3 nm mn mm nn nv
4 Db BD Bd Bb BB

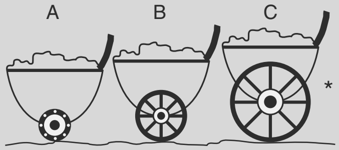
Xy Yy YX XX XY   nn mn nv nm mm
○ ○ ● ○ ○   ○ ○ ○ ● ○
g6 Gg 6g G6 6G   BD BB Bd Db Bb
○ ○ ○ ○ ●   ● ○ ○ ○ ○

Mechanical Reasoning (25 min., 60 items)

Understanding basic mechanical principles of machinery, tools, and motion is important for occupations such as carpentry, mechanics, engineering, and machine operation.

SAMPLE ITEM

Which load will be easier to pull through soft sand?

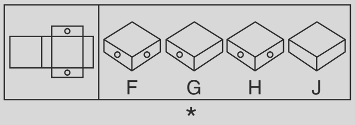


Space Relations (25 min., 50 items)

Measures the ability to visualize a three-dimensional object from a two-dimensional pattern, and to visualize how this object would look if rotated in space; important in drafting, architecture, design, carpentry, and dentistry.

SAMPLE ITEM

Choose the one figure that can be made from the pattern.



Spelling (10 min., 40 items)

Measures one's ability to spell common English words; a useful skill in many academic and vocational pursuits.

SAMPLE ITEM

Decide which word is not spelled correctly in the group below.

*A paragraf
B dramatic
C circular
D audience

Language Usage (15 min., 40 items)

Measures the ability to detect errors in grammar, punctuation, and capitalization; needed in most jobs requiring a college degree.

SAMPLE ITEM

Decide which of the four parts of the sentence below contains an error. If there is no error, mark the space on your answer sheet for the letter next to No Error.

Jane and Tom/ is going/ to the office/ this morning.
A *B C D
E No Error

Source: Adapted from *Differential Aptitude Tests: Fifth Edition.* Copyright © 1990 by Pearson Education, Inc. and/or its affiliates. Reproduced with permission. All rights reserved.

of skills and abilities you expect each student to have acquired before entering the new instruction.

Admissions Testing

Multiple-Assessment for Admission Test scores, previous grades, letters of recommendation, interviews, and biographical information on out-of-school accomplishments are among kinds of information colleges and selective high schools use to make admissions decisions. Some private and parochial high schools, for example, use a battery of achievement and aptitude tests to screen applicants. Testing sessions are usually held at the local high school, and its staff generally administers the tests.

College Admissions Tests Two widely used college admissions testing programs are the College Entrance Examination Board's *SAT Reasoning Test* and the *ACT Assessment Program* published by American College Testing. Both programs administer secure tests. Both administer the tests through local testing centers (usually high schools and colleges) on preestablished dates several times during the year. For both programs, test booklets and answer sheets are returned to their respective publishers for scoring, recording, and processing results to the colleges the students designate.

SAT Reasoning Test The College Entrance Examination Board (CEEB), currently located in New York City, was formed around 1899 to help select colleges in the northeastern United States coordinate their admissions testing requirements. The tests developed out of that effort around 1926 and were created by Carl Brigham, associate secretary for the CEEB (Donlon & Angoff, 1971). Educational Testing Service in Princeton, New Jersey, currently develops the test for the CEEB. More than 2 million college candidates take the test each year.

The program includes the *SAT Reasoning Test* and *SAT Subject Tests*. Students generally take one or both types during their junior or senior year of high school. Only the *SAT Reasoning Test* is discussed in this chapter. It has three parts: a Critical Reading section, a Mathematics section, and a Writing section. The Critical Reading section emphasizes reading and word knowledge. Two types of multiple-choice items are used, with questions based on short and long reading passages. (Analogy questions, a feature of past *SAT*s, have been eliminated.) The Mathematics section emphasizes quantitative thinking using arithmetic, algebra (both Algebra I and II content), and geometry knowledge. It includes items on topics such as estimation, exponential growth, absolute value, functional notation, linear functions, manipulations with exponents, and tangent lines. Two types of items are used: standard multiple-choice and student-produced response. For student-produced response items, examinees "bubble in" their numerical answers on a special answer sheet (there are no choices). The Writing section contains both multiple-choice grammar (identifying error in sentences, and improving sentences and paragraphs) items and a 20-minute written essay. In the essay, candidates are asked to develop a point of view on an issue, and are evaluated on their ability to reason and use evidence to support their ideas. The test booklets and answer sheets are sent to the Educational Testing Service (ETS) for scoring, recording, and processing scores to the colleges the student designates. A student's essay responses are posted on a Website that college admission officers can access and read.

ACT Information Program The *ACT Assessment Program* was formed in 1959, with the ideas and help of E. F. Lindquist, among others. This admissions testing program was originally conceived to be of a different character than the *SAT* program. Whereas initially CEEB was concerned primarily with the private select colleges of the Northeast, the ACT program initially sought to serve midwestern public colleges and universities. What these colleges needed was help in (a) eliminating the few incapable students who were applying, (b) providing guidance services for those admitted, and (c) measuring broad educational development rather than narrower verbal and quantitative aptitudes (Lindquist in Feister & Whitney, 1968). Today American College Testing in Iowa City is as active a research and test development enterprise as is the Educational Testing Service. More than 1.4 million college candidates take the test each year.

The *ACT Assessment Program* has four components: *Tests of Educational Development*, Course/Grade Information, Student Profile, and *ACT Interest Inventory*. The *Tests of Educational Development* comprise four multiple-choice subtests: English, Mathematics, Reading, and Science Reasoning. In addition, there is an optional writing test. The English Test emphasizes standard written English conventions and rhetorical skills. The Mathematics Test emphasizes quantitative reasoning and problem solving using prealgebra, algebra, geometry, and trigonometry knowledge. The Reading Test contains passages representative of topics in social studies, natural sciences, fiction, and humanities. The items focus on using inference and reasoning for reading comprehension. The Science Reasoning Test contains several sets of related data tables, diagrams, and verbal descriptions drawn from biology, chemistry, physics, and earth/space science. The items focus on interpreting data, interpreting experimental results, and reasoning with respect to alternative viewpoints. The developers view the test items as "work samples"—simulations of the kinds of learning activities typically required of the first-year college student.

The Course/Grade Information section asks candidates to self-report their grades in 30 courses from what is usually included in a college preparatory curriculum in English, mathematics, natural sciences, social studies, language, and the arts. The Student Profile section asks candidates to report 200 pieces of information including educational plans, interests, and needs; special educational needs, interests, and goals; college extracurricular plans; financial aid; high school extracurricular activities; out-of-class accomplishments; and so on. Among other purposes, this questionnaire permits the student to indicate any special talents and accomplishments not reflected in course grades (such as winning a state debate). The *ACT Interest Inventory* consists of a list of activities, and students indicate whether they would like, dislike, or are indifferent about doing each activity on the list. The interest inventory helps students get a better idea of how their career interests fit into the mainstream of various major areas of college.

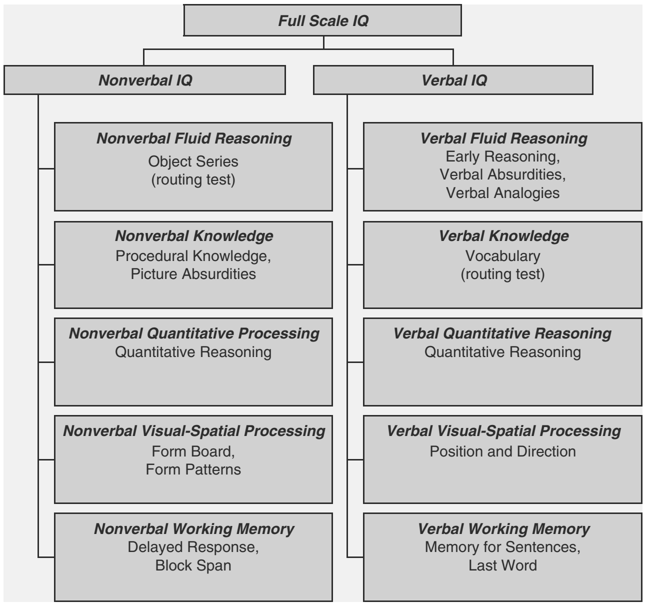
INDIVIDUALLY ADMINISTERED TESTS OF GENERAL SCHOLASTIC APTITUDES

Stanford-Binet Intelligence Scale

History The *Stanford-Binet Intelligence Scale* is a widely used, individually administered test of general scholastic aptitude. First prepared in 1916 by Lewis M. Terman as a translation and revision of the *Binet-Simon Scale,* the test was revised in 1937 (with Maud A. Merrill), revised again in 1960, renormed in 1972, revised and renormed in 1986, and revised and renormed for the 2003 (fifth edition by Gale Roid). See Becker's (2003) *History of the* Stanford-Binet Intelligence Scales.

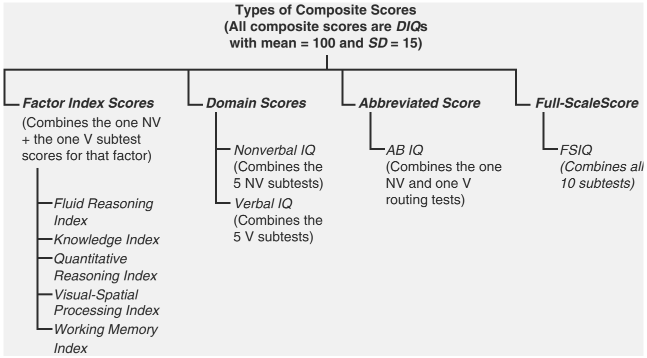
Content The *Stanford-Binet V* is used with a wide range of ages, from 2 years old through adults age 85+. You can gain an idea of the nature and content of this assessment instrument by studying the diagram that follows. The diagram shows that the subtests are clustered into five nonverbal areas (factors) and five verbal areas (factors). The scores from the five nonverbal factors are combined to obtain the Nonverbal IQ; the scores from the five verbal factors are combined to obtain the Verbal IQ.

Structure of the Stanford-Binet V



Not all items in each subtest are administered because within each subtest items are arranged in order of increasing difficulty. The object series and vocabulary subtest is given first and is used as a routing test. The student's performance on this test, along with the student's age or estimated ability, tells the psychologist the difficulty level on the other tests at which he or she should begin testing the student. If a quick (and less reliable) estimate of the Full Scale IQ is desired, the psychologist can stop after administering the verbal and nonverbal routing tests. The standard scores on these two tests are combined to obtain an Abbreviated Full Scale IQ score.

Scores A student's raw score on each subtest is converted to a normalized standard score called a **standard age score (*SAS*)** for the subtest. The *SAS*s for each of the 10 subtests have a mean of 10 and standard deviation of 3 in the norm group having the same age as the student being tested. The 10 *SAS*-scores from the subtests are combined in different ways to make 9 different composite scores. There are four kinds of composite scores: Factor Index Scores, Domain Scores, Abbreviated Score, and Full-Scale Score. All the composite scores are deviation IQs (*DIQ*s) with a mean of 100 and a standard deviation of 15, as explained in Chapter 16, Equation 16.4. Within each composite score type, there are from one to five different *DIQ*-scores. The diagram below shows how these four composite scores are formed.


Types of Composite Scores
(All composite scores are DIQs
with mean = 100 and SD = 15)
Factor Index Scores
(Combines the one NV + the one V subtest scores for that factor)
Fluid Reasoning Index
Knowledge Index
Quantitative Reasoning Index
Visual-Spatial Processing Index
Working Memory Index
Domain Scores
Nonverbal IQ (Combines the 5 NV subtests)
Verbal IQ (Combines the 5 V subtests)
Abbreviated Score
AB IQ (Combines the one NV and one V routing tests)
Full-ScaleScore
FSIQ (Combines all 10 subtests)

The subtest scores and the different composite scores are used by school psychologists or counseling psychologists along with other information (e.g., school records, other test results, interviews, and reports from teachers and parents) for (a) describing a profile of a student's intellectual skills and abilities, (b) classifying a student into a diagnostic category (e.g., attention deficit disorder/ hyperactivity disorder), or (c) placing a student into a special educational program (e.g., a gifted program).

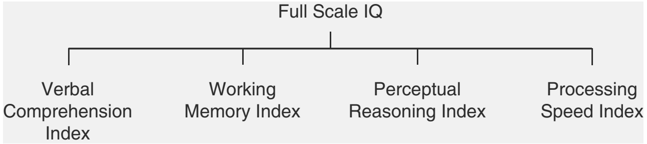
The concepts of **mental age** and intelligence quotient (IQ) are no longer used with tests such as the *Stanford-Binet V* (or any other modern intelligence test). Thus, the ratio IQ (= 100 times mental age divided by chronological age) has been replaced by the *DIQ*.

Norm-Referenced Character Tests of scholastic aptitude describe a student's ability as the student's location in a norm group having the same age as the student. If the student's intellectual development does not keep pace with others in the norm group, the student will receive a lower *DIQ*. Norms become outdated and from time to time a test will have to be renormed.

Wechsler Intelligence Scales

Another widely used set of individual tests is the *Wechsler Intelligence Scales*. This set consists of three different intelligence tests, each designed for use with a different age level: (a) *Wechsler Preschool and Primary Scale of Intelligence–III* (*WPPSI–III*), 2 years, 6 months to 7 years, 3 months; (b) *Wechsler Intelligence Scale for Children–IV* (*WISC–IV*), 6 to 16 years, 11 months; and (c) *Wechsler Adult Intelligence Scale–III* (*WAIS–IV*), 16 to 90 years.

General Design All the Wechsler tests have a similar general design, although the items are not identical. The items are organized into subtests. The items within a subtest are similar in content but differ in difficulty. (The subtests on the different scales—*WPPSI, WISC,* and *WAIS*—contain different types of items, however.) Subtests are clustered into groups to represent different factors or aspects of general ability. These form a hierarchical pattern of abilities as shown below, but with some slight differences for the different age-level tests:



The four factors at the lowest level of the hierarchy are described as follows (Pearson, 2008):

Verbal Comprehension: One's ability to listen to, understand, and give spoken responses to verbal questions. It includes skills in understanding information presented verbally, using words to think and reason, and using words to express thoughts.

Working Memory: One's ability to learn new information and retain it in memory as one completes a task. It includes skills in paying attention, in concentrating, and in mental reasoning.

Perceptual Reasoning: One's ability to examine and think about pictures and designs, and solve problems without using words. It involves skills working quickly with visual information to solve nonverbal problems.

Processing Speed: One's ability to scan symbols and make judgments about them quickly. It involves skills in paying attention, hand-eye coordination, and mental problem solving.

Scores All the Wechsler scales report subtest results as normalized, standard scores (mean = 10, standard deviation = 3). All of the scales report the total or Full Scale IQ as *DIQ*-scores (mean = 100, standard deviation = 15). The four factor indexes are also *DIQ*-scores with a mean of 100 and a standard deviation of 15. The norm group to which a student is referenced is the group of students with the same age.

WAIS–IV The *Wechsler Adult Intelligence Scale–IV* is used with ages 16 to 90 years. It overlaps with the *WISC—IV* for age 16. It contains 15 subtests, but only 10 are used to calculate the four indexes as follows (supplemental subtests are in parentheses):

Verbal Comprehension Index: Vocabulary, Similarities, Information, (Comprehension)

Perceptual Reasoning Index: Block Design, Matrix Reasoning, Visual Puzzles, (Figure Weights), (Picture Completion)

Working Memory Index: Digit Span, Arithmetic, (Letter-Number Sequencing)

Processing Speed Index: Symbol Search, Coding, (Cancellation)

The *WAIS–IV* is generally considered to be a reasonably valid and reliable tool for assessing general cognitive ability. Following are some of its limitations (Sattler, 1992): (a) It does not provide low enough scores for persons with severe intellectual disabilities, (b) it does not provide high enough scores for persons with extremely gifted mental ability, and (c) the range of subtest scaled scores is restricted for some age groups.

WISC–IV The *Wechsler Intelligence Scale for Children—IV* is used with children ages 6 years through 16 years, 11 months. It contains 10 core and 5 supplemental subtests. In the Verbal Comprehension category are Similarities, Vocabulary, and Comprehension (Information and Word Reasoning are supplemental subtests). In the Perceptual Reasoning category are Block Design, Picture Concepts, and Matrix Reasoning (Picture Completion is a supplemental subtest). In the Working Memory category are Digit Span and Letter-Number Sequencing (Arithmetic is a supplemental subtest). In the Processing Speed category is Coding and Symbol Search (Cancellation is a supplemental subtest).

The *WISC–IV* was standardized with the *Wechsler Individual Achievement Test–II* (*WIAT–II*), an individually administered basic skills achievement test. This makes it possible to compare mental ability results from the *WISC–IV* with the *WIAT–II,* a comparison often made for students experiencing learning difficulties in school. This assists in the process of establishing individualized education programs (IEPs).

The *WISC–IV* is generally considered to be a good test of overall mental ability. Among its strengths (Sattler, 1992) are its (a) high-quality norms; (b) good reliability and validity; (c) usefulness in diagnosing cognitive abilities of most students; (d) good features of the materials, administration, and scoring; and (e) extensive research literature. Among its limitations (Sattler, 1992) are its (a) lack of usefulness for extremely low- and high-ability children, (b) restriction of the range of scores for certain subtests and age levels, (c) lack of appropriate norms when a subtest is substituted, (d) susceptibility to large practice effects on the Performance Scale, and (e) potential for penalizing students who do not place a premium on speed of responding. Like the *Stanford-Binet V,* the *WISC–IV* is used by psychologists, along with other information, for developing students' profiles of intellectual skills and abilities, classification in diagnostic categories, or placement in special educational programs. The *SB–V* and *WISC–IV* are different tests; you should not expect them to come to the exact same conclusion about a student.

WPPSI–III The *Wechsler Preschool and Primary Scale of Intelligence–III* is used with children ages 2 years, 6 months years through 7 years, 3 months. It overlaps with the *WISC–IV* for ages 6 years through 7 years, 3 months. For this overlapping age range, the *WISC–IV* is recommended (Sattler, 1992). The *WPPSI–III* contains 15 subtests, 8 of which are supplemental subtests. In the Verbal category are Information, Vocabulary, and Word Reasoning (Comprehension and Similarities are supplemental subtests). In the Performance category are Block Design, Matrix Reasoning, and Picture Concepts (Picture Completion and Object Assembly are supplemental subtests). The Processing Speed or visual-motor category contains Coding and Symbol Search. The Full Scale IQ uses scores from the three Verbal, the three Performance, and one of the Processing Speed subtests.

Kaufman Assessment Battery for Children

General Description The *Kaufman Assessment Battery for Children–II* (*KABC–II*) is an individually administered test of general intelligence

(Kaufman & Kaufman, 2004). It is used with children ages 3 through 18. The *KABC–II* differs in several ways from other approaches to measuring scholastic aptitude described thus far: (a) The subtests were derived from a differential psychological model (Cattell-Horn-Carroll [CHC] model; see Alfonso, Flanagan, & Radwan, 2005) and neuropsychological theory (Luria model; see Das, 2002); (b) a psychologist must decide before testing which one of the two interpretive models to use with a particular child and base the overall score only on the chosen model; and (c) there is a deliberate attempt to organize the testing to make it "fairer" to students not in the mainstream culture and for certain students with language-affected disabilities. In the norming sample, nonmainstream ethnic groups had average scores that were slightly higher when the Luria model was used than when the CHC model was used. The use of different models for defining cognitive ability is helpful, too, when professionals are developing IEPs for students.

Content The *KABC–II* is organized into five scales. The scales and their subtests are organized as follows. The names in brackets are the scale names when the CHC model is used.

- *Sequential Processing Scale [Short-Term Memory]* (G_{sm})[1]—One's ability to remember an ordered series of images or ideas and use this memory to do a task. Requires repeating a sequence of numbers or identifying a sequence of pictures that the examiner says. Includes two subtests (and one supplemental subtest): Number Recall and Word Order (Hand Movements).
- *Simultaneous Processing Scale [Visual Processing]* (G_v)—One's ability to consider an array of information and process the parts of the array simultaneously to do the task. This form of thinking requires the student to visualize and integrate the elements in the array presented, so it is called visual processing ability. Consists of six subtests (and one supplemental subtest): Face Recognition, Triangles, Conceptual Thinking, Pattern Recognition, Rover, and Block Counting (Gestalt Closure), plus Pattern Reasoning and Story Completion for ages 5 and 6.
- *Planning Ability Scale [Fluid Reasoning]* (G_f)—One's ability to understand a nonverbal problem, generate a hypothesis about how to solve it, test that solution, and revise it if necessary. Students must use verbal reasoning to solve the nonverbal problems. Includes two subtests: Pattern Reasoning and Story Completion, for ages 7 to 18.
- *Learning Ability Scale [Long-Term Storage and Retrieval]* (G_{lr})—One's ability to successfully complete different types of tasks that require learning something new. Some tasks require immediate recall of the newly learned information and others require using that information after a period of delay. Includes two subtests: Atlantis and Rebus.
- *Knowledge Scale [Crystallized Ability]* (G_c)—One's ability to express knowledge and understanding of objects and events in the mainstream culture. Students are asked to express their knowledge of words and facts, when questions are asked verbally and through pictorial stimuli. They respond either verbally or by pointing. Consists of three subtests: Riddles, Expressive Vocabulary, and Verbal Knowledge.

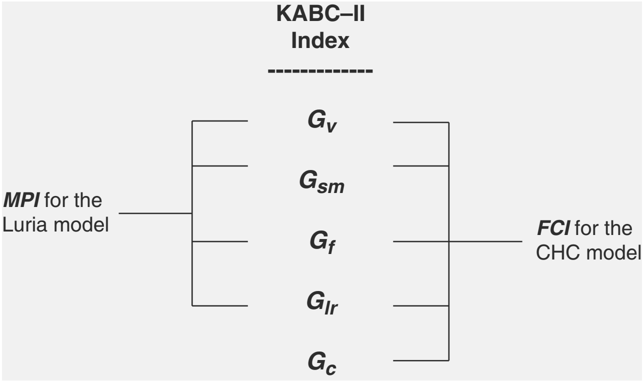
A readable explanation of the Luria model and the CHC model, a full description of the *KABC–II*, and the history of the *KABC–II* are found in Kaufman, Lichtenberger, Fletcher-Janzen, and Kaufman (2005).

Scores The *KABC–II* provides *DIQ*-scores (mean = 100, standard deviation = 15) for each of the five scales: Sequential Processing, Simultaneous Processing, Planning Ability, Learning Ability, and Knowledge. Each of the subtests within these scales is reported as a normalized standard score (mean = 10, standard deviation = 3). In addition, there are three *DIQ* composite scores: the *Mental Processing Index* (*MPI*), the *Fluid-Crystallized Index* (*FCI*), and the *Non-Verbal Index* (*NVI*). Any one student can be assigned only *MPI* and *NVI* or *FCI* and *NVI*.

You will recall that the examiner must choose to use either the Luria model or the CHC model before testing a student. If the examiner chooses the Luria model, then the student can receive the *MPI* composite but not the *FCI* composite; if the examiner chooses the CHC model, the student receives the *FCI* composite, not the *MPI*. The difference is that the *MPI* does not include the Knowledge/

[1] The notation used is *G* with a subscript. The *G* represents "general ability factor," first postulated by Spearman (1927). This factor can be decomposed into subfactors like the ones defined here for the CHC model. The subscript on the *G* denotes the subfactor.

Crystallized Ability (G_c) subtest because it is not administered under the Luria model. Here is the structure:



Usefulness of the *KABC–II* Approach The authors suggest the following uses for the *FCI* and *MPI* (Kaufman et al., 2005):

- The *FCI* (i.e., all five areas) should be used for the majority of students; when there is a suspicion of a reading, written expression, or mathematics disability; for a child with mental retardation; for a child with attention-deficit/hyperactivity disorder (ADHD); for a child with an emotional or behavioral disorder; and with a child who is gifted.
- The *MPI* (i.e., exclude administration of the Knowledge component) should be used with children from bilingual backgrounds; children from nonmainstream cultural backgrounds whose verbal development is problematic; and with children who have language disorders, autism, or deafness/hearing loss.

Until the research on the *KABC–II* has been completed we cannot properly evaluate these proposed uses.

ASSESSING ADAPTIVE BEHAVIOR

Meaning of Adaptive Behavior

Tests such as the *Stanford-Binet*, the *WISC*, and the *KABC* measure general scholastic ability. A school setting, of course, is not the only environment in which persons are expected to cope. Some students may appear to teachers and other school personnel to suffer from intellectual disabilities, but their families, neighbors, and peers accept the students and consider the students normal in all other facets of life. It is recommended, therefore, that before labeling a student as having intellectual disabilities, the student's ability to cope with the demands of his or her environment outside classroom learning be assessed. According to the American Association on Intellectual and Developmental Disabilities (2002), "*Intellectual disability* is a disability characterized by significant limitations both in intellectual functioning and in *adaptive behavior*, which covers many everyday social and practical skills. This disability originates before the age of 18." [emphasis added].

Adaptive behavior assessment focuses on how independently students can care for themselves and how well they can cope with the demands placed on them by the immediate culture in which they are living. Thus, these types of assessments focus on a student's success as a family member, consumer, wage earner, member of a nonacademic peer group, person interacting with adults, and person caring for his or her health and physical needs—that is, skills in the three domains of conceptual, social, and practical adaptive skills. A psychological report for a student often includes assessment of the student's adaptive behavior as well as his or her general scholastic aptitude. This section presents one example.

Vineland Adaptive Behavior Scales (VABS)

The *Vineland Adaptive Behavior Scales–II* (*VABS–II*; revised and renormed by Sparrow, Cicchetti, & Balla, 2005) is a developmental checklist that assesses adaptive behavior in five areas: Communication (expressive, receptive, and written); Daily Living Skills (domestic tasks, personal habits, behavior outside the home); Socialization (interpersonal relations, play and leisure skills, coping); Motor Skills (gross and fine motor); and Maladaptive Behavior (inappropriate and undesirable behavior). The first three areas are assessed for persons from birth through 90 years (and low-functioning adults). Motor Skills assessment is limited to children younger than 9 years and the Maladaptive Behavior area to children 5 years and older. A trained interviewer completes the assessment by interviewing a child's parent or caregiver.

The *VABS–II* has four editions: (1) Interview Edition, Survey Form, which provides standard scores for each of the five areas as well as a total adaptive behavior score (ages birth to 90); (2) Interview Edition, Expanded Form, which in addition to the standard scores provides specific, detailed information for preparing educational and habilitation programs (ages birth to 90);

(3) Classroom Edition, for ages 3 to 22, which a teacher completes as a questionnaire and which provides standard scores for four adaptive behavior areas as well as a total adaptive behavior score; and (4) Parent/Caregiver Rating forms. A qualified professional is needed to interpret the scores. The *VABS–II* provides national norms for all editions. For the two interview editions, special supplemental norms are available for adults with intellectual disabilities (residential and nonresidential), children with hearing impairments, children with visual impairments, and children with emotional disturbances (the latter three groups in residential settings).

According to the authors, the *VABS–II* may be used for diagnosing adaptive behavior deficits, determining eligibility for special services, planning intervention programs, and tracking progress in development. The authors have planned for its use with populations of individuals with intellectual disabilities, autism spectrum disorders (ASDs), ADHD, posttraumatic brain injury, hearing impairment, and dementia/Alzheimer's disease.

ASSESSING VOCATIONAL AND CAREER INTERESTS

What Are Interests?

Attitudes, Interests, and Values Three characteristics of students are closely related: attitudes, interests, and values. Questionnaires very often assess these characteristics. The questionnaires appear similar because, when responding to the questionnaire, a student reads several statements and expresses his or her degree of agreement with the statements.

In spite of their similarity, the three concepts are not identical. To interpret assessment results properly, you must distinguish among these three concepts. An **attitude** is a positive or negative feeling about a physical object, a type of people, a particular person, a government or other social institution's policy, ideas, or the like. For example, when a student expresses agreement with the statement, "My mathematics class helps me become a better person," the student is expressing his attitude toward the mathematics class.

Interests, on the other hand, are preferences for specific types of activities when a person is not under external pressure. For example, when a student expresses agreement with the statement, "I enjoy working on the mathematics problems my teacher assigns," the student is expressing her interest in a mathematics activity.

Values, unlike attitudes and interests, are long-lasting beliefs of the importance of certain life goals, a lifestyle, a way of acting, or a way of life. For example, when a student expresses agreement with the statement, "I consider it more important to be one of the best students in mathematics than to be one of the best players in a football game," the student is expressing his valuing of mathematics success over football success.

When studying ways of assessing attitudes, interests, and values, keep in mind that the methods you use for assessing them are highly susceptible to students' providing socially desirable responses, as opposed to frank personal responses. Therefore, questionnaires can assess only what an individual wishes to reveal.

Focus on Career Interests This cluster of interests is important as students begin to prepare themselves for further schooling and for the world of work. No single piece of information is sufficient for a student to use in making vocational decisions, of course. However, the student's interest in various activities associated with specific types of work or work environments is an important consideration. Besides knowing the duty requirements of the job market, and his own abilities and aptitudes, a student should also understand his own interests regarding work-related activities. Thus, the types of career interest inventories described next can provide one source of information to help a student make educational and vocational choices.

Expressed, Manifested, Tested, and Inventoried Interests

Interest inventories of the type described in this section are limited to only vocational interests or career interests; and career interests are narrowed even further. You may find it useful to distinguish among expressed, manifested, tested, and inventoried interests.

Expressed Interests **Expressed interests** are obtained when you ask students directly about their interests. The interests a student verbally professes when you ask the student directly may not express her true preferences: A youngster may express an interest in being a doctor, for example, because she perceives it as something parents expect. Or a teenager may say she wants to be a rock musician just to see the reaction of her parents.

Manifested Interests **Manifested interests** are inferred from what a student actually does, or the activities in which the student actually participates. When you attempt to infer students' interests from their activities, you may misjudge. For example, you may conclude a boy is interested in athletics because he participates in the junior high track team, but you later find out he only wants to be with his friends after school.

Tested Interests **Tested interests** are those you infer from the results of assessing a student's information and knowledge of a particular subject matter. For example, you may hypothesize that a student who has a lot of scientific knowledge and information has more interest in science than a student who knows little about science. Such knowledge assessments are not used very often in current vocational counseling practice.

Inventoried Interests **Inventoried interests** are identified through various paper-and-pencil tests or interest inventories. A limitation here is that the interests you discover through a particular interest inventory do not represent all interests or even all career interests. Further, as with other forms of educational and psychological assessment, the interest patterns identified with one publisher's interest inventory may not be the same ones that could be identified with others. When counseling students, you should use all three interests—expressed, manifested, and inventoried—to assess a student's interest patterns.

Vocational Interest Inventories

Vocational interest inventories are formal paper-and-pencil questionnaires that help students express their likes and dislikes about a very wide range of work and other activities. A student's pattern of interests is then determined from these responses. This profile or pattern of interests becomes one source of information a student can use for career exploration, counseling, and decision making.

Building Interest Inventories The traditional rationale for describing a person's inventoried interests has been called the **people-similarity rationale** (Cole & Hanson, 1975, p. 6): "If a person likes the same things that people in a particular job like, the person will be satisfied with the job."

Certain parts of the *Kuder Occupational Interest Survey* (*KOIS*) and the *Strong Interest Inventory* (*SII*) follow this rationale. Both the *KOIS* and the Occupational Scales of the *SII*, for example, are **empirically keyed scales**, made up of items especially selected because research has shown that responses to these items clearly differentiate between the persons who are currently and happily employed in a particular occupation and people in general.

A second rationale has been called the **activity-similarity rationale** (Cole & Hanson, 1975, p. 6): "If people like activities similar to the activities required by a job, they will like those job activities and consequently be satisfied with their job."

Inventories built using this rationale present the students with lists of activities that are similar to those required of persons working in certain jobs or studying certain subjects. The developers assume that if a person has an identifiable pattern of likes and dislikes common to a particular job, that person will be satisfied with that job. Among the inventories developed using this rationale are the *ACT Interest Inventory* and the *Ohio Vocational Interest Survey, Second Edition* (*OVIS–II*).

Formats of Interest Items Figure 18.5 shows sample items from two vocational interest inventories. Notice that the items from one inventory, the *ACT,* ask students to rate each activity or statement on a like–dislike continuum. Items from the other inventory, the *KOIS,* present activities in sets of three (triads). These latter items ask the student to mark the one activity in the triad that the student most ("M") likes and the one activity the student least ("L") likes. This is equivalent to asking a student to rank the three activities from most liked to least liked. This approach, called a **forced-choice item format**, was designed to overcome the tendency for some students to have very high personal standards for "like," whereas others have very low standards. When this difference in standards occurs, two students who may in fact have the *same order* of likes or dislikes for an activity may mark their answer sheets differently. Measurement experts have criticized the forced-choice format, however, because using it results in a statistical

FIGURE 18.5 Examples of items on two interest process inventories.

Sources: The test items from the *ACT Interest Inventory*, p. 10 in "Registering for the ACT Assessment." Copyright by ACT, Inc. Reproduced by permission of the publisher. The items from the *Kuder Occupational Interest Survey, Form DD* are reproduced by permission. Published by National Career Assessment Services Inc.™, PO Box 277, Adel, IA 50003. 800-314-8972/www.kuder.com. Copyright © by National Career Assessment Services, Inc™. All rights reserved.

The *ACT Interest Inventory*

I would *dislike* doing this activity	D
I am *indifferent* (don't care one way or the other)	I
I would *like* doing this activity	L

1. Explore a science museum
2. Play jazz in a combo
3. Help settle an argument between friends

Kuder Occupational Interest Survey, Form DD

* 1.	Visit an art gallery	M	L
	Browse in a library	M	L
	Visit a museum	M	L
2.	Collect autographs	M	L
	Collect coins	M	L
	Collect stones	M	L

Notes:

*M=most liked, L=least liked

artifact that causes a negative correlation among the scales of the inventory.

Item Content The pioneers of the interest inventory technique used a variety of content to survey interests, including asking examinees likes and dislikes of job titles, school subject matter, hobbies, leisure activities, work activities, types of persons, and type of reading material, and assessing examinees' personal characteristics (Davis, 1980). Over time, however, the concept of interests narrowed to the world of work and careers. Today the content of most inventories is limited exclusively to lists of activities, and most are concerned with work activities. An exception is the *SII,* which uses a large variety of content to measure a person's interest in relation to persons working in a wide range of careers.

Strong Interest Inventory

A number of vocational interest inventories are listed in Appendix K. One of the inventories, the *Strong Interest Inventory* (*SII*), is briefly described here.

Organization The 2004 revision, which is used with persons 14 years old and older, is composed of 291 items. A person is presented with occupations, subject areas, activities, leisure activities, and people and rates each item on a 5-point scale: strongly like, like, indifferent, dislike, or strongly dislike. The responses are then scored via a computer (tests cannot be scored locally) and reported as standard scores on various scales. The *SII* results are reported to the examinee in four ways: (1) 6 General Occupational Themes, which describe a person's overall pattern of occupational interests; (2) 30 Basic Interest Scales, which describe the somewhat narrower categories of interest areas a person likes within the 6 General Occupational Themes; (3) 211 Occupational Scales, which describe the extent to which a person's likes and dislikes are similar to persons working in specific occupations; and (4) 5 Personal Style Scales.

General Occupational Themes The student receives a score on each of six areas, which were adapted from Holland (1973). These are Realistic, Conventional, Investigative, Enterprising, Artistic, and Social. A counselor can use the student's profile regarding these areas to help her understand her overall or general pattern of interests, work activities, personal values, and how she appears to be oriented to the world of work. A normalized standard score (*T*-score) is reported for each of the six themes.

Basic Interest Scales The 30 scales are grouped in clusters under each of the six themes. For example, six of the Basic Interest Scales—Teaching and Education, Social Sciences, Human Resources and Training, Healthcare Services, Religion and Spirituality, and Counseling and Helping—are clustered under the general occupational theme of Social. The Basic Interest Scales are intermediate between the General Occupational Themes and the more specific Occupational Scales (described next). The Basic Interest Scales describe the clusters of interests a student has. These activity areas may be common to several specific occupations (for example, there are many types of teachers).

Occupational Scales Paralegal, respiratory specialist, florist, artist, and social science teacher are some of the occupational scales. The occupational scales are organized under each of the six general occupation themes. A person receives a standard score for each occupation based on the combined male and female norms. This score describes how similar the student's pattern of likes and dislikes is to persons who are experienced and satisfied in each occupation. In the report to the student within each General Occupational Theme, occupations are ordered according to the student's similarity to persons of their own gender. Occupations listed first are those for which the students' *SII* responses are most similar to persons of their own gender who are working in that occupation.

Special Scales There are five Personal Style Scales describing how the student approaches learning, people, and the workplace: Work Style, Learning Environment, Leadership Style, Risk Taking, and Team Orientation. In addition to these scales, there is a special administrative index, the Typicality Index. This index is used to assess the consistency of a student's responses to items that are very highly correlated in the norm group in an attempt to detect random and atypical response patterns. A counselor uses this index to help decide whether a student responded well enough to make the results meaningful. If not, the counselor will need to explore with the student individually why he did not respond consistently to very similar items.

Male-Female Differences Although the *SII* has a single booklet for both males and females, there are 122 Occupational Scales for men and 122 Occupational Scales for women for a total of 244 Occupational Scales. The authors have kept scale reporting separate for each gender because (a) there are large differences in the strength of interests in the two genders in many areas and (b) combined-sex (unisex) scales appear less valid for many occupations. Students may understand their interests better if they can compare them to both like-gender and opposite-gender norms. This may be especially helpful to students who are thinking of entering occupations dominated by the gender opposite of their own.

ASSESSING ATTITUDES

Attitudes and Their Characteristics

Attitudes are characteristics of persons that describe their positive and negative feelings toward particular objects, situations, institutions, persons, or ideas. Keep in mind that attitudes are learned, and once learned they direct or guide the students' actions.The attitudes of older students and adults are changeable, but it is much easier to change the attitudes of younger students. You cannot observe students' attitudes directly; you must infer them from the students' actions or from responses to an attitude questionnaire. Because students can fake their responses to attitude questionnaires, you should interpret the results very cautiously.

Attitudes differ in both **direction** and **intensity**. Two students may hold the same positive attitude (direction), but the students may differ greatly regarding the strength of feeling (intensity) they attach to that attitude. Students' attitudes will also differ in **affective saliency** or emotionality. Two students may have the same positive attitude, but one may become much more emotional than the other regarding it.

ASSESSING PERSONALITY DIMENSIONS

A variety of techniques have been developed to measure various aspects of personality. A person using a personality test must be trained in psychological interpretation of the results. This usually requires extensive graduate work in counseling or school psychology and a lengthy supervised internship. Teachers will encounter the results of such tests, however, if they are part of a child study team or if they read psychological reports of students. Thus, some familiarity with a few basic concepts of personality measurement is in order.

Assessment Approaches

The kind of personality tests a counselor or psychologist uses depends primarily on the psychological orientation of the particular professional. Currently, there is no standard model or conception of personality, nor do counselors and psychologists agree on which particular aspects of personality are most important to assess. The kinds of personality tests used in psychological reports a teacher may encounter depend on the background and training of the psychologist assigned to a given student's case.

Projective Techniques

Projective Hypothesis Two broad methods for assessing personality dimensions are projective techniques and structured techniques. **Projective personality test techniques** present the examinee with ambiguous stimuli and ask the examinee to respond to them. The proponents of this technique assume that an examinee's interpretations of these vague stimuli will reveal the examinee's innermost needs, feelings, and conflicts, even though the examinee is unaware of what he or she is revealing. This assumption is known as the **projective hypothesis** (Frank, 1939). A trained examiner is needed to interpret an examinee's responses. Examples of projective personality tests are the *Rorschach Test*, the *Thematic Apperception Test* (*TAT*), word association tests, various sentence-completion tests, certain picture arrangement tests, and various figure drawing tests. School psychologists use projective tests less often now than in the past.

Sentence-Completion Assessments Sentence **completion tests of personality** ask the examinee to complete sentences related to various aspects of self and interpersonal relations (e.g., "Compared with most families, mine . . ."). The results of content analyses are used similarly to generate hypotheses about a subject's personality.

Structured Techniques

Definition **Structured personality assessment techniques** follow very specific rules for administering, scoring, and interpreting the tests. Usually they follow a response-choice format: yes-no, true-false, or multiple-choice. Examples of structured personality tests are the *Guilford-Zimmerman Temperament Survey*, the *Minnesota Multiphasic Personality Inventory–Second Edition*, the *California Psychological Inventory*, and the *Personality Inventory for Children–Second Edition*.

Self-Report Characteristic Each test is sometimes referred to as a **self-report personality inventory** because it requires examinees to respond to the items in a way that describes personal feelings. For instance, examinees may be asked whether the statement, "I usually express my personal opinions to others," is true of themselves.

Dimensions of Personality Another characteristic of structured personality inventories is that the items are related to various scales or personality dimensions. The *Guilford-Zimmerman Temperament Survey*, for example, reports an examinee's profile with respect to 10 scales: general activity, restraint, ascendance (leadership), sociability, emotional stability, objectivity, friendliness, thoughtfulness, personal relations, and masculinity.

Usefulness of Tests for the Teacher

Self-report personality inventories require persons to (Cunningham, Thorndike, & Hagen, 1991) (a) read and comprehend each item, (b) be able to understand their own actions enough to know whether a given statement is true of them, and (c) be willing to respond honestly and frankly. Reading in the context of personality testing requires that students understand the items well enough to be able to decide the degree to which the statements apply to their own lives. To decide whether a statement applies, a student must view that behavior objectively, which may not be within the repertoire of a poorly adjusted student. Finally, if a student is neither able nor willing to respond frankly to the items, a distorted personality description may result. This lack of frankness may occur more often when testing children who feel vulnerable or threatened if they reveal their feelings to the teacher or, more generally, to the school. Considering the shortcomings of self-report personality and adjustment inventories, some measurement experts conclude they have a limited role in education.

CONCLUSION

There is a lot more to say about scholastic aptitude, career interests, attitudes, and personality tests than we have been able to fit into one chapter. We hope, however, that this chapter has been a useful introduction to some types of tests you will encounter as you work in schools.

This textbook has been about educational assessment of all sorts. Returning to the theme that began in Chapter 1, the goal of good educational assessment is to provide valid—and you now know in some detail what that means—information to support sound educational decisions. Educational assessment happens in individual sessions (as for some of the aptitude tests described in this chapter), in classrooms, and at the school, district, state, national, and international levels. Education-related decisions are made at all these levels, as well. We hope that, armed with understandings from this textbook and experience from your course- and schoolwork, you are prepared to participate in sound assessment of students and in the resulting educational decisions.

EXERCISES

1. Describe several school situations in which it is less helpful to know a student's level of specific skill development than to know a student's general intellectual skill development in setting expectations for learning new material.
2. Read each of the following statements of educational needs. For each statement choose a test that possibly could meet the stated need. Choose from among the tests in this set to respond to this exercise: *OLSAT, DAT,* readiness tests, *SAT, ACT,* aptitude tests for a specific subject, *SB–V, WISC–IV, KABC–II*.
   a. "In addition to finding out a student's verbal and quantitative aptitudes, I would like to know how well the student processes symbols and other nonverbal material."
   b. "I'd like to give all ninth graders a test that would provide information helpful to them in making career decisions."
   c. "I would like to know which of my fifth-grade students could learn computer programming quickly and well."
   d. "I need a general ability test for a student who recently arrived from Cuba."
   e. "I need a college admissions test that gives me information that I can use in guidance and counseling activities as well as in admissions."
3. Both the *DAT* and *Strong Interest Inventory* report results on separate gender norms. Explain why they do so, and discuss whether this practice is helpful to the career and further schooling planning of females.
4. Read the following statements. Each statement expresses a student's status with respect to achievement, aptitude, attitude, interest, or values. Classify each statement into one of the five categories.
   a. "I am in control of my learning in this class at all times."
   b. "I like science fiction stories better than biographies."
   c. "I think my math class is boring."
   d. "It is more important for me to be in personal control of my working hours than to earn a high salary."
   e. "I am constantly striving to be the best student in this school."
5. Visit your school's guidance department and determine how its counselors use interest inventories. Share your findings with your classmates.

Educational Assessment Knowledge and Skills for Teachers

I. Teachers should understand learning in the content area they teach.

The primary purpose for most educational assessment is to advance student learning. In order to be able to assess students well and to make sound decisions based on the results, teachers must understand general principles about how students learn, understand deeply the content area(s) they teach, and understand specific learning progressions within a content area. Selecting and communicating clear learning targets (II and III below), designing or selecting assessments that evaluate them (IV and V below), and knowing the difference between a domain of learning and the assignments and assessments selected to embody it in the classroom depends on these understandings. Interpreting student work directly and interpreting scores derived from student performance on assessments (VI and VII below) require a solid undertanding of both the content area and how learning in that content area typically proceeds. Decisions about what to do in light of a teacher's interpretations of assessment results (VIII below) similarly depend on understanding of typical learning progressions in the content area.

II. Teachers should be able to articulate clear learning objectives that are congruent with both the content and depth of thinking implied by standards and curriculum goals, in such a way that the objectives are attainable and assessable.

The basis for instruction and assessment is the competence to define and describe the knowledge and skills students need to learn in clear, attainable, and assessable ways. These "know and be able to do" statements focus teachers' instructional planning and students' intentions for learning. To support effective learning, learning objectives must be sound, coherent with standards and curriculum goals, and clearly communicated to students. They must be objectives that the students can achieve. They must be assessable so that both students and teacher will know whether and to what degree they have been achieved.

III. Teachers should have a repertoire of strategies for communicating what achievement of a learning target looks like.

Once articulated, learning targets need to be shared and communicated with students, and often parents and colleagues, as well. Teachers should have a repertoire of several strategies in each of several different communication modes—telling, showing, and having students discover—for communicating learning targets in the content areas they teach.

IV. Teachers should understand and be skilled in using the range of assessment options available and the purposes and uses of each.

Teachers should have both knowledge of the various kinds of test item formats and performance tasks and the skills to create sound, appropriate assessments from them. To do this, teachers need to understand the concept that a standard or learning objective is a domain and an assessment samples from that domain. To create a sound assessment of a domain of learning, teachers need to understand the concept of validity (including reliability) as the degree to which assessment information supports its intended purpose and use, and they need the skills to prepare assessments that yield valid results. Teachers should understand issues of fairness and issues of accessibility (including available accommodations and modifications for students with disabilities, and their implications for validity, accessibility,

Source: From *Educational Assessment Knowledge and Skills for Teachers*, by Susan M. Brookhart. Unpublished manuscript.

From Appendix A of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

and fairness), as these are related to valid assessment outcomes.

V. Teachers should have the skills to analyze classroom questions, test items, and performance assessment tasks to ascertain the specific knowledge and thinking skills required for students to do them.

Teachers should be able to apply these skills as they

a. Ask their own classroom questions or write their own test items and performance tasks.

b. Evaluate questions in teachers' manuals, other curriculum material, and prepared test items and performance tasks (e.g., from textbook materials, or from other teachers) for potential use.

c. Provide feedback directly on student work.

d. Use assessment results to plan future instruction.

e. Coach students to analyze their own assessment results. A hugely important teacher job is to facilitate students being able to articulate learning objectives, assess and interpret their own work, and use this information for future study and performance.

VI. Teachers should be able to construct scoring schemes that quantify student performance on classroom assessments into useful information for decisions about students, classrooms, schools, and districts. These decisions should lead to improved student learning, growth, or development.

Teachers need quantitative knowledge and reasoning skills for use with both classroom and large-scale assessments (VI and VII in this outline). For classroom assessments, teachers should know and be able to use various methods of scoring individual items or tasks (right/wrong for items or checklists and multipoint methods including rubrics and rating scales) in the classroom. They should know and be able to use accurately various methods of aggregating scores into meaningful wholes (points, percents, grades). Their understanding should include the basics of simple linear scaling, weighting components, and precision of the results. Sound quantitative reasoning should lead to scores that can serve as dependable evidence about a student's classroom learning and be used in such a way that improved learning results.

VII. Teachers should be able to administer external assessments and interpret their results for decisions about students, classrooms, schools, and districts. These decisions should lead to improved student learning, growth, and development.

Teachers should know how to administer state- or district-mandated standardized assessments, or school-mandated common assessments like end-of-course exams or common final exams, according to standardized directions, and understand why such standardization is necessary for interpreting these assessments' results. Teachers should be able to interpret conventional norm- and criterion-referenced scores reported on external test results, including but not limited to: understanding measurement error and confidence intervals; limiting generalization to the construct assessed and not beyond; understanding the difference between grade-equivalent scores and grade-level instructional objectives; and understanding differences between scores for individual students and class- or school-level aggregated scores. Teachers should be able to use these understandings about score meaning to improve students' learning.

VIII. Teachers should be able to articulate their interpretations of assessment results and their reasoning about the educational decisions based on assessment results to the educational populations they serve (students and their families, class, school, community).

Teachers should be able to apply this skill as they

a. Speak understandably with students about the results of their own assessments and what that means for the next steps in improving their learning.

b. Speak understandably with parents about the results of their children's classroom assessments, report card grades, and external standardized assessments, the decisions made or recommended on the basis of these assessments, and the intended consequences and follow-up.

c. Participate productively in discussions with parents and guidance counselors, and sometimes students, regarding decisions about student guidance or placement (including work on IEPs) , and implementation and follow-up of those decisions.

d. Participate productively and in informed ways in committee or school-wide discussions about assessment-related issues, including but not limited to: curriculum materials adoption and/or curriculum reform, report card reform, grading policies, accountability policies and reporting, program or school evaluation, and teacher evaluation.

IX. Teachers should understand and carry out their legal and ethical responsibilities in assessment as they conduct their work.

Understandings and commitments to legal and ethical responsibilities should be evident in all the work a teacher does. Areas of understanding include, but are not limited to, test preparation, confidentiality of information, opportunity to learn, and due process. Teachers should make decisions based on results from multiple, appropriate assessments (for example, in the areas of grading policies or drawing conclusions from external test results).

Code of Fair Testing Practices in Education (Revised)

Prepared by the Joint Committee on Testing Practices

The *Code of Fair Testing Practices in Education (Code)* is a guide for professionals in fulfilling their obligation to provide and use tests that are fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics. Fairness is a primary consideration in all aspects of testing. Careful standardization of tests and administration conditions helps to ensure that all test takers are given a comparable opportunity to demonstrate what they know and how they can perform in the area being tested. Fairness implies that every test taker has the opportunity to prepare for the test and is informed about the general nature and content of the test, as appropriate to the purpose of the test. Fairness also extends to the accurate reporting of individual and group test results. Fairness is not an isolated concept, but must be considered in all aspects of the testing process.

The *Code* applies broadly to testing in education (admissions, educational assessment, educational diagnosis, and student placement) regardless of the mode of presentation, so it is relevant to conventional paper-and-pencil tests, computer-based tests, and performance tests. It is not designed to cover employment testing, licensure or certification testing, or other types of testing outside the field of education. The *Code* is directed primarily at professionally developed tests used in formally administered testing programs. Although the *Code* is not intended to cover tests made by teachers for use in their own classrooms, teachers are encouraged to use the guidelines to help improve their testing practices.

The *Code* addresses the roles of test developers and test users separately. Test developers are people and organizations that construct tests, as well as those that set policies for testing programs. Test users are people and agencies that select tests, administer tests, commission test development services, or make decisions on the basis of test scores. Test developer and test user roles may overlap, for example, when a state or local education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of the test scores.

Many of the statements in the *Code* refer to the selection and use of existing tests. When a new test is developed, when an existing test is modified, or when the administration of a test is modified, the *Code* is intended to provide guidance for this process.

The *Code* is not intended to be mandatory, exhaustive, or definitive, and may not be applicable to every situation. Instead, the *Code* is intended to be aspirational, and is not intended to take precedence over the judgment of those who have competence in the subjects addressed.

The *Code* provides guidance separately for test developers and test users in four critical areas:

A. Developing and Selecting Appropriate Tests

B. Administering and Scoring Tests

C. Reporting and Interpreting Test Results

D. Informing Test Takers

Source: Copyright 2004 by the Joint Committee on Testing Practices. This material may be reproduced in whole or in part without fees or permission, provided that acknowledgment is made to the Joint Committee on Testing Practices. Reproduction and dissemination of this document are encouraged. This edition replaces the first edition of the *Code,* which was published in 1988. Please cite this document as follows: *Code of Fair Testing Practices in Education (Revised)* (2004). Washington, DC: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, Science Directorate, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; http://www.apa.org/science/jctpweb.html.) Contact APA for additional copies.

From Appendix B of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

The *Code* is intended to be consistent with the relevant parts of the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999). The *Code* is not meant to add new principles over and above those in the Standards or to change their meaning. Rather, the *Code* is intended to represent the spirit of selected portions of the Standards in a way that is relevant and meaningful to developers and users of tests, as well as to test takers and/or their parents or guardians. States, districts, schools, organizations, and individual professionals are encouraged to commit themselves to fairness in testing and safeguarding the rights of test takers. The *Code* is intended to assist in carrying out such commitments.

The *Code* has been prepared by the Joint Committee on Testing Practices, a cooperative effort among several professional organizations. The aim of the Joint Committee is to act, in the public interest, to advance the quality of testing practices. Members of the Joint Committee include the American Counseling Association (ACA), the American Educational Research Association (AERA), the American Psychological Association (APA), the American Speech-Language-Hearing Association (ASHA), the National Association of School Psychologists (NASP), the National Association of Test Directors (NATD), and the National Council on Measurement in Education (NCME).

A. DEVELOPING AND SELECTING APPROPRIATE TESTS*

Test Developers

Test developers should provide the information and supporting evidence that test users need to select appropriate tests.

1. Provide evidence of what the test measures, the recommended uses, the intended test takers, and the strengths and limitations of the test, including the level of precision of the test scores.
2. Describe how the content and skills to be tested were selected and how the tests were developed.
3. Communicate information about a test's characteristics at a level of detail appropriate to the intended test users.
4. Provide guidance on the levels of skills, knowledge, and training necessary for appropriate review, selection, and administration of tests.
5. Provide evidence that the technical quality, including reliability and validity, of the test meets its intended purposes.
6. Provide to qualified test users representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports.
7. Avoid potentially offensive content or language when developing test questions and related materials.
8. Make appropriately modified forms of tests or administration procedures available for test takers with disabilities who need special accommodations.
9. Obtain and provide evidence on the performance of test takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed.

Test Users

Test users should select tests that meet the intended purpose and that are appropriate for the intended test takers.

1. Define the purpose for testing, the content and skills to be tested, and the intended test takers. Select and use the most appropriate test based on a thorough review of available information.
2. Review and select tests based on the appropriateness of test content, skills tested, and content coverage for the intended purpose of testing.
3. Review materials provided by test developers and select tests for which clear, accurate, and complete information is provided.
4. Select tests through a process that includes persons with appropriate knowledge, skills, and training.
5. Evaluate evidence of the technical quality of the test provided by the test developer and any independent reviewers.
6. Evaluate representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports before selecting a test.
7. Evaluate procedures and materials used by test developers, as well as the resulting test, to ensure that potentially offensive content or language is avoided.
8. Select tests with appropriately modified forms or administration procedures for test takers with disabilities who need special accommodations.
9. Evaluate the available evidence on the performance of test takers of diverse subgroups. Determine to the extent feasible which performance differences may

* Many of the statements in the *Code* refer to the selection of existing tests. However, in customized testing programs test developers are engaged to construct new tests. In those situations, the test development process should be designed to help ensure that the completed tests will be in compliance with the *Code*.

have been caused by factors unrelated to the skills being assessed.

B. ADMINISTERING AND SCORING TESTS

Test Developers

Test developers should explain how to administer and score tests correctly and fairly.

1. Provide clear descriptions of detailed procedures for administering tests in a standardized manner.
2. Provide guidelines on reasonable procedures for assessing persons with disabilities who need special accommodations or those with diverse linguistic backgrounds.
3. Provide information to test takers or test users on test question formats and procedures for answering test questions, including information on the use of any needed materials and equipment.
4. Establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting.
5. Provide procedures, materials and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, provide adequate training for scorers.
6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.
7. Develop and implement procedures for ensuring the confidentiality of scores.

Test Users

Test users should administer and score tests correctly and fairly.

1. Follow established procedures for administering tests in a standardized manner.
2. Provide and document appropriate procedures for test takers with disabilities who need special accommodations or those with diverse linguistic backgrounds. Some accommodations may be required by law or regulation.
3. Provide test takers with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing.
4. Protect the security of test materials, including respecting copyrights and eliminating opportunities for test takers to obtain scores by fraudulent means.
5. If test scoring is the responsibility of the test user, provide adequate training to scorers and ensure and monitor the accuracy of the scoring process.
6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.
7. Develop and implement procedures for ensuring the confidentiality of scores.

C. REPORTING AND INTERPRETING TEST RESULTS

Test Developers

Test developers should report test results accurately and provide information to help test users interpret test results correctly.

1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.
2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.
3. Specify appropriate uses of test results and warn test users of potential misuses.
4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.
5. Encourage test users to base decisions about test takers on multiple sources of appropriate information, not on a single test score.
6. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.
7. Provide test results in a timely fashion and in a manner that is understood by the test taker.
8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.

Test Users

Test users should report and interpret test results accurately and clearly.

1. Interpret the meaning of the test results, taking into account the nature of the content, norms or comparison groups, other technical evidence, and benefits and limitations of test results.
2. Interpret test results from modified test or test administration procedures in view of the impact those modifications may have had on test results.

3. Avoid using tests for purposes other than those recommended by the test developer unless there is evidence to support the intended use or interpretation.
4. Review the procedures for setting performance standards or passing scores. Avoid using stigmatizing labels.
5. Avoid using a single test score as the sole determinant of decisions about test takers. Interpret test scores in conjunction with other information about individuals.
6. State the intended interpretation and use of test results for groups of test takers. Avoid grouping test results for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use. Report procedures that were followed in determining who were and who were not included in the groups being compared and describe factors that might influence the interpretation of results.
7. Communicate test results in a timely fashion and in a manner that is understood by the test taker.
8. Develop and implement procedures for monitoring test use, including consistency with the intended purposes of the test.

D. INFORMING TEST TAKERS

Test Developers or Test Users

Under some circumstances, test developers have direct communication with the test takers and/or control of the tests, testing process, and test results. In other circumstances the test users have these responsibilities.

Test developers or test users should inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores.

1. Inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers.
2. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether a test should be taken—including indications of any consequences that may result from not taking the test (e.g., not being eligible to compete for a particular scholarship)——and whether there is an available alternative to the test.
3. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid.
4. Provide test takers or their parents/guardians with information about responsibilities test takers have, such as being aware of the intended purpose and uses of the test, performing at capacity, following directions, and not disclosing test items or interfering with other test takers.
5. Inform test takers or their parents/guardians how long scores will be kept on file and indicate to whom, under what circumstances, and in what manner test scores and related information will or will not be released. Protect test scores from unauthorized release and access.
6. Describe procedures for investigating and resolving circumstances that might result in canceling or withholding scores, such as failure to adhere to specified testing procedures.
7. Describe procedures that test takers, parents/guardians, and other interested parties may use to obtain more information about the test, register complaints, and have problems resolved.

Note: The membership of the working group that developed the *Code of Fair Testing Practices in Education* and of the Joint Committee on Testing Practices that guided the working group is as follows: Peter Behuniak, PhD; Lloyd Bond, PhD; Gwyneth M. Boodoo, Phd; Wayne Camara, PhD; Ray Fenton, PhD; John J. Fremer, PhD (Cochair); Sharon M. Goldsmith, PhD; Bert F. Green, PhD; William G. Harris, PhD; Janet E. Helms, PhD; Stephanie H. McConaughy, PhD; Julie P. Noble, PhD; Wayne M. Patience, PhD; Carole L. Perlman, PhD; Douglas K. Smith, PhD; Janet E. Wall, EdD (Cochair); Pat Nellor Wickwire, PhD; Mary Yakimowski, PhD. Lara Frumkin, PhD, of the APA served as staff liaison. The Joint Committee intends that the *Code* be consistent with and supportive of existing codes of conduct and standards of other professional groups who use tests in educational contexts. Of particular note are the Responsibilities of Users of Standardized Tests (Association for Assessment in Counseling, 1989), APA Test User Qualifications (2000), ASHA Code of Ethics (2001), Ethical Principles of Psychologists and Code of Conduct (1992), NASP Professional Conduct Manual (2000), NCME Code of Professional Responsibility (1995), and Rights and Responsibilities of Test Takers: Guidelines and Expectations (Joint Committee on Testing Practices, 2000).

Code of Professional Responsibilities in Educational Measurement

Prepared by the NCME Ad Hoc Committee on the Development of a Code of Ethics: Cynthia B. Schmeiser, ACT—Chair; Kurt F. Geisinger, State University of New York; Sharon Johnson-Lewis, Detroit Public Schools; Edward D. Roeber, Council of Chief State School Officers; William D. Schafer, University of Maryland

PREAMBLE AND GENERAL RESPONSIBILITIES

As an organization dedicated to the improvement of measurement and evaluation practice in education, the National Council on Measurement in Education (NCME) has adopted this Code to promote professionally responsible practice in educational measurement. Professionally responsible practice is conduct that arises from either the professional standards of the field, general ethical principles, or both.

The purpose of the Code of Professional Responsibilities in Educational Measurement, hereinafter referred to as the Code, is to guide the conduct of NCME members who are involved in any type of assessment activity in education. NCME is also providing this Code as a public service for all individuals who are engaged in educational assessment activities in the hope that these activities will be conducted in a professionally responsible manner. Persons who engage in these activities include local educators such as classroom teachers, principals, and superintendents; professionals such as school psychologists and counselors; state and national technical, legislative, and policy staff in education; staff of research, evaluation, and testing organizations; providers of test preparation services; college and university faculty and administrators; and professionals in business and industry who design and implement educational and training programs.

This Code applies to any type of assessment that occurs as part of the educational process, including formal and informal, traditional and alternative techniques for gathering information used in making educational decisions at all levels. These techniques include, but are not limited to, large-scale assessments at the school, district, state, national, and international levels; standardized tests; observational measures; teacher-conducted assessments; assessment support materials; and other achievement, aptitude, interest, and personality measures used in and for education.

Although NCME is promulgating this Code for its members, it strongly encourages other organizations and individuals who engage in educational assessment activities to endorse and abide by the responsibilities relevant to their professions. Because the Code pertains only to uses of assessment in education, it is recognized that uses of assessments outside of educational contexts, such as for employment, certification, or licensure, may involve additional professional responsibilities beyond those detailed in this Code.

The Code is intended to serve an educational function: to inform and remind those involved in educational assessment of their obligations to uphold the integrity of the manner in which assessments are developed, used, evaluated, and marketed. Moreover, it is expected that the Code will stimulate thoughtful discussion of what constitutes professionally responsible assessment practice at all levels in education.

SECTION 1: RESPONSIBILITIES OF THOSE WHO DEVELOP ASSESSMENT PRODUCTS AND SERVICES

Those who develop assessment products and services, such as classroom teachers and other assessment specialists, have a professional responsibility to strive to produce assessments that are of the highest quality. Persons who develop assessments have a professional responsibility to:

1.1 ensure that assessment products and services are developed to meet applicable professional, technical, and legal standards.

Source: © 1995 by the National Council on Measurement in Education. Reprinted by permission of the publisher. Available: http://www.natd.org/Code_of_Professional_Responsibilities.html

From Appendix C of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

1.2 develop assessment products and services that are as free as possible from bias due to characteristics irrelevant to the construct being measured, such as gender, ethnicity, race, socioeconomic status, disability, religion, age, or national origin.

1.3 plan accommodations for groups of test takers with disabilities and other special needs when developing assessments.

1.4 disclose to appropriate parties any actual or potential conflicts of interest that might influence the developers' judgment or performance.

1.5 use copyrighted materials in assessment products and services in accordance with state and federal law.

1.6 make information available to appropriate persons about the steps taken to develop and score the assessment, including up-to-date information used to support the reliability, validity, scoring and reporting processes, and other relevant characteristics of the assessment.

1.7 protect the rights to privacy of those who are assessed as part of the assessment development process.

1.8 caution users, in clear and prominent language, against the most likely misinterpretations and misuses of data that arise out of the assessment development process.

1.9 avoid false or unsubstantiated claims in test preparation and program support materials and services about an assessment or its use and interpretation.

1.10 correct any substantive inaccuracies in assessments or their support materials as soon as feasible.

1.11 develop score reports and support materials that promote the understanding of assessment results.

SECTION 2: RESPONSIBILITIES OF THOSE WHO MARKET AND SELL ASSESSMENT PRODUCTS AND SERVICES

The marketing of assessment products and services, such as tests and other instruments, scoring services, test preparation services, consulting, and test interpretive services, should be based on information that is accurate, complete, and relevant to those considering their use. Persons who market and sell assessment products and services have a professional responsibility to:

2.1 provide accurate information to potential purchasers about assessment products and services and their recommended uses and limitations.

2.2 not knowingly withhold relevant information about assessment products and services that might affect an appropriate selection decision.

2.3 base all claims about assessment products and services on valid interpretations of publicly available information.

2.4 allow qualified users equal opportunity to purchase assessment products and services.

2.5 establish reasonable fees for assessment products and services.

2.6 communicate to potential users, in advance of any purchase or use, all applicable fees associated with assessment products and services.

2.7 strive to ensure that no individuals are denied access to opportunities because of their inability to pay the fees for assessment products and services.

2.8 establish criteria for the sale of assessment products and services, such as limiting the sale of assessment products and services to those individuals who are qualified for recommended uses and from whom proper uses and interpretations are anticipated.

2.9 inform potential users of known inappropriate uses of assessment products and services and provide recommendations about how to avoid such misuses.

2.10 maintain a current understanding about assessment products and services and their appropriate uses in education.

2.11 release information implying endorsement by users of assessment products and services only with the users' permission.

2.12 avoid making claims that assessment products and services have been endorsed by another organization unless an official endorsement has been obtained.

2.13 avoid marketing test preparation products and services that may cause individuals to receive scores that misrepresent their actual levels of attainment.

SECTION 3: RESPONSIBILITIES OF THOSE WHO SELECT ASSESSMENT PRODUCTS AND SERVICES

Those who select assessment products and services for use in educational settings, or help others do so, have important professional responsibilities to make sure that the assessments are appropriate for their intended use. Persons who select assessment products and services have a professional responsibility to:

3.1 conduct a thorough review and evaluation of available assessment strategies and instruments that might be valid for the intended uses.

3.2 recommend and/or select assessments based on publicly available documented evidence of their technical quality and utility rather than on unsubstantiated claims or statements.

3.3 disclose any associations or affiliations that they have with the authors, test publishers, or others involved with the assessments under consideration for purchase and refrain from participation if such associations might affect the objectivity of the selection process.

3.4 inform decision makers and prospective users of the appropriateness of the assessment for the intended uses, likely consequences of use, protection of examinee rights, relative costs, materials and services needed to conduct or use the assessment, and known limitations of the assessment, including potential misuses and misinterpretations of assessment information.

3.5 recommend against the use of any prospective assessment that is likely to be administered, scored, and used in an invalid manner for members of various groups in our society for reasons of race, ethnicity, gender, age, disability, language background, socioeconomic status, religion, or national origin.

3.6 comply with all security precautions that may accompany assessments being reviewed.

3.7 immediately disclose any attempts by others to exert undue influence on the assessment selection process.

3.8 avoid recommending, purchasing, or using test preparation products and services that may cause individuals to receive scores that misrepresent their actual levels of attainment.

SECTION 4: RESPONSIBILITIES OF THOSE WHO ADMINISTER ASSESSMENTS

Those who prepare individuals to take assessments and those who are directly or indirectly involved in the administration of assessments as part of the educational process, including teachers, administrators, and assessment personnel, have an important role in making sure that the assessments are administered in a fair and accurate manner. Persons who prepare others for, and those who administer, assessments have a professional responsibility to:

4.1 inform the examinees about the assessment prior to its administration, including its purposes, uses, and consequences; how the assessment information will be judged or scored; how the results will be kept on file; who will have access to the results; how the results will be distributed; and examinees' rights before, during, and after the assessment.

4.2 administer only those assessments for which they are qualified by education, training, licensure, or certification.

4.3 take appropriate security precautions before, during, and after the administration of the assessment.

4.4 understand the procedures needed to administer the assessment prior to administration.

4.5 administer standardized assessments according to prescribed procedures and conditions and notify appropriate persons if any nonstandard or delimiting conditions occur.

4.6 not exclude any eligible student from the assessment.

4.7 avoid any conditions in the conduct of the assessment that might invalidate the results.

4.8 provide for and document all reasonable and allowable accommodations for the administration of the assessment to persons with disabilities or special needs.

4.9 provide reasonable opportunities for individuals to ask questions about the assessment procedures or directions prior to and at prescribed times during the administration of the assessment.

4.10 protect the rights to privacy and due process of those who are assessed.

4.11 avoid actions or conditions that would permit or encourage individuals or groups to receive scores that misrepresent their actual levels of attainment.

SECTION 5: RESPONSIBILITIES OF THOSE WHO SCORE ASSESSMENTS

The scoring of educational assessments should be conducted properly and efficiently so that the results are reported accurately and in a timely manner. Persons who score and prepare reports of assessments have a professional responsibility to:

5.1 provide complete and accurate information to users about how the assessment is scored, such as the reporting schedule, scoring process to be used, rationale for the scoring approach, technical characteristics, quality control procedures, reporting formats, and the fees, if any, for these services.

5.2 ensure the accuracy of the assessment results by conducting reasonable quality control procedures before, during, and after scoring.

5.3 minimize the effect on scoring of factors irrelevant to the purposes of the assessment.

5.4 inform users promptly of any deviation in the planned scoring and reporting service or schedule and negotiate a solution with users.

5.5 provide corrected score results to the examinee or the client as quickly as practicable should errors be found that may affect the inferences made on the basis of the scores.

5.6 protect the confidentiality of information that identifies individuals as prescribed by state and federal law.

5.7 release summary results of the assessment only to those persons entitled to such information by state or federal law or those who are designated by the party contracting for the scoring services.

5.8 establish, where feasible, a fair and reasonable process for appeal and rescoring the assessment.

SECTION 6: RESPONSIBILITIES OF THOSE WHO INTERPRET, USE, AND COMMUNICATE ASSESSMENT RESULTS

The interpretation, use, and communication of assessment results should promote valid inferences and minimize invalid ones. Persons who interpret, use, and communicate assessment results have a professional responsibility to:

6.1 conduct these activities in an informed, objective, and fair manner within the context of the assessment's limitations and with an understanding of the potential consequences of use.

6.2 provide to those who receive assessment results information about the assessment, its purposes, its limitations, and its uses necessary for the proper interpretation of the results.

6.3 provide to those who receive score reports an understandable written description of all reported scores, including proper interpretations and likely misinterpretations.

6.4 communicate to appropriate audiences the results of the assessment in an understandable and timely manner, including proper interpretations and likely misinterpretations.

6.5 evaluate and communicate the adequacy and appropriateness of any norms or standards used in the interpretation of assessment results.

6.6 inform parties involved in the assessment process how assessment results may affect them.

6.7 use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions.

6.8 avoid making, and actively discourage others from making, inaccurate reports, unsubstantiated claims, inappropriate interpretations, or otherwise false and misleading statements about assessment results.

6.9 disclose to examinees and others whether and how long the results of the assessment will be kept on file, procedures for appeal and rescoring, rights examinees and others have to the assessment information, and how those rights may be exercised.

6.10 report any apparent misuses of assessment information to those responsible for the assessment process.

6.11 protect the rights to privacy of individuals and institutions involved in the assessment process.

SECTION 7: RESPONSIBILITIES OF THOSE WHO EDUCATE OTHERS ABOUT ASSESSMENT

The process of educating others about educational assessment, whether as part of higher education, professional development, public policy discussions, or job training, should prepare individuals to understand and engage in sound measurement practice and to become discerning users of tests and test results. Persons who educate or inform others about assessment have a professional responsibility to:

7.1 remain competent and current in the areas in which they teach and reflect that in their instruction.

7.2 provide fair and balanced perspectives when teaching about assessment.

7.3 differentiate clearly between expressions of opinion and substantiated knowledge when educating others about any specific assessment method, product, or service.

7.4 disclose any financial interests that might be perceived to influence the evaluation of a particular assessment product or service that is the subject of instruction.

7.5 avoid administering any assessment that is not part of the evaluation of student performance in a course if the administration of that assessment is likely to harm any student.

7.6 avoid using or reporting the results of any assessment that is not part of the evaluation of student performance in a course if the use or reporting of results is likely to harm any student.

7.7 protect all secure assessments and materials used in the instructional process.

7.8 model responsible assessment practice and help those receiving instruction to learn about their professional responsibilities in educational measurement.

7.9 provide fair and balanced perspectives on assessment issues being discussed by policymakers, parents, and other citizens.

SECTION 8: RESPONSIBILITIES OF THOSE WHO EVALUATE EDUCATIONAL PROGRAMS AND CONDUCT RESEARCH ON ASSESSMENTS

Conducting research on or about assessments or educational programs is a key activity in helping to improve the understanding and use of assessments and educational programs. Persons who engage in the evaluation of educational programs or conduct research on assessments have a professional responsibility to:

8.1 conduct evaluation and research activities in an informed, objective, and fair manner.

8.2 disclose any associations that they have with authors, test publishers, or others involved with the assessment and refrain from participation if such associations might affect the objectivity of the research or evaluation.

8.3 preserve the security of all assessments throughout the research process as appropriate.

8.4 take appropriate steps to minimize potential sources of invalidity in the research and disclose known factors that may bias the results of the study.

8.5 present the results of research, both intended and unintended, in a fair, complete, and objective manner.

8.6 attribute completely and appropriately the work and ideas of others.

8.7 qualify the conclusions of the research within the limitations of the study.

8.8 use multiple sources of relevant information in conducting evaluation and research activities whenever possible.

8.9 comply with applicable standards for protecting the rights of participants in an evaluation or research study, including the rights to privacy and informed consent.

AFTERWORD

As stated at the outset, the purpose of the *Code of Professional Responsibilities in Educational Measurement* is to serve as a guide to the conduct of NCME members who are engaged in any type of assessment activity in education. Given the broad scope of the field of educational assessment as well as the variety of activities in which professionals may engage, it is unlikely that any code will cover the professional responsibilities involved in every situation or activity in which assessment is used in education. Ultimately, it is hoped that this Code will serve as the basis for ongoing discussions about what constitutes professionally responsible practice. Moreover, these discussions will undoubtedly identify areas of practice that need further analysis and clarification in subsequent editions of the Code. To the extent that these discussions occur, the Code will have served its purpose.

Summaries of Taxonomies of Educational Objectives: Cognitive, Affective, and Psychomotor Domains

FIGURE D.1 **Categories and subcategories of the Bloom et al. taxonomy of cognitive objectives.**

- 1.00 **Knowledge**
  - 1.10 **Knowledge of Specifics**
  - 1.11 **Knowledge of Terminology** Knowledge of the referents for specific symbols (verbal and nonverbal). . . .
  - 1.12 **Knowledge of Specific Facts** Knowledge of dates, events, persons, places, etc.
  - 1.20 **Knowledge of Ways and Means of Dealing with Specifics**
  - 1.21 **Knowledge of Conventions** Knowledge of characteristic ways of treating and presenting ideas and phenomena.
  - 1.22 **Knowledge of Trends and Sequences** Knowledge of the processes, directions, and movements of phenomena with respect to time.
  - 1.23 **Knowledge of Classifications and Categories** Knowledge of the classes, sets, divisions, and arrangements that are regarded as fundamental for a given subject field, purpose, argument, or problem.
  - 1.24 **Knowledge of Criteria** Knowledge of the criteria by which facts, principles, and conduct are tested or judged.
  - 1.25 **Knowledge of Methodology** Knowledge of the methods of inquiry, techniques, and procedures employed in a particular subject field as well as those employed in investigating particular problems and phenomena.
- 2.00 **Comprehension**
  - 2.10 **Translation** Comprehension as evidenced by the care and accuracy with which the communication is paraphrased or rendered from one language or form of communication to another.
  - 2.20 **Interpretation** The explanation or summarization of a communication.
  - 2.30 **Extrapolation** The extension of trends or tendencies beyond the given data to determine implications, consequences, corollaries, effects, etc., that are in accordance with the conditions described in the original communication.
- 3.00 **Application** The use of abstractions in particular and concrete situations. The abstractions may be in the form of general ideas, rules of procedures, or generalized methods.
- 4.00 **Analysis**
  - 4.10 **Analysis of Elements** Identification of the elements included in a communication.
  - 4.20 **Analysis of Relationships** The connections and interactions between elements and parts of a communication.
  - 4.30 **Analysis of Organized Principles** The organization, systematic arrangement, and structure that hold the communication together.
- 5.00 **Synthesis**
  - 5.10 **Production of a Unique Communication** The development of a communication in which the writer or speaker attempts to convey ideas, feelings, and/or experiences to others.
  - 5.20 **Production of a Plan or Proposed Set of Operations** The development of a plan of work or the proposal of a plan of operations.
  - 5.30 **Derivation of a Set of Abstract Relations** The development of a set of abstract relations either to classify or to explain particular data or phenomena, or the deduction of propositions and relations from a set of basic propositions or symbolic representations.
- 6.00 **Evaluation**
  - 6.10 **Judgments in Terms of Internal Evidence** Evaluation of the accuracy of a communication from such evidence as logical accuracy, consistency, and other internal criteria.
  - 6.20 **Judgments in Terms of External Criteria** Evaluation of material with reference to selected or remembered criteria.

Source: Adapted from *Taxonomy of Educational Objectives Book 1. Cognitive Domain* (pp. 201–207), edited by Benjamin S. Bloom et al. Published by Allyn and Bacon, Boston, MA. Copyright © 1984 by Pearson Education. Reprinted with permission of the publisher.

From Appendix D of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FIGURE D.2 Gagné's levels of complexity in human skills, characteristics of responses to tasks assessing these capacities, and examples of specific objectives written for each capacity.

Type of ability or capacity	Characteristics of responses to assessment tasks	Example of a specific learning target
1. **Discrimination:** ability to respond appropriately to stimuli that differ. The stimuli can differ in one or more physical attributes such as size, shape, (capacity verb: *discriminates*)	The learner's response must indicate that the learner has distinguished between the different stimuli. The leader may do this by indicating "same" or "different."	Given learner two cardboard cutouts, one a triangle shape and the other a square shape, the learner can point to the one that is a "square."
2. **Concrete concept:** ability to identify a stimulus as belonging to a particular class or category. The members of the class have one or more physical properties in common. (capacity verb: *identifies*)	The learner's response must indicate that two or more members of the class have been identified.	Given several differently shaped figures of various colors and shapes, half of which have triangular shapes, the learner can point to all the "triangles."
3. **Defined concept:** ability to demonstrate what is meant by a defined class of objects, events, or relations—that is, demonstrate an understanding of a concept. (capacity verb: *classifies*)	The learner's response must go beyond memorization to identify specific instances of the defined concept and to show how these instances are related to each other (and are thereby members of the same concept or category).	Given descriptions and brief biographies of each of several different persons not born in this country, the learner is able to identify all the persons who are immigrants and state their relationship to each other.
4. **Rule:** ability to make responses that indicate a rule is being applied in a variety of different situations. (capacity verb: *demonstrates*)	The learner's response must indicate that a particular rule is being applied in one or more concrete instances, but the learner need not be able to state the rule.	Given a "story" problem of the type presented in class involving two single-digit addends, the pupil is able to add the digits correctly.
5. **Higher-order rule:** (problem solving): ability to form a new (for the learner) rule to solve a problem, by combining two or more previously learned rules. (capacity verb: *generates*)	The learner's response must indicate that a new complex rule has been "invented" and applied to solve a problem that is new or novel for the learner. Once the rule is invented, the learner should be able to apply it to other situations (transfer of learning).	Given an announcement about a specific job opening for which the learner is qualified, the learner is able to generate and write an appropriate letter of application for that job.
6. **Cognitive strategies:** ability to use internal processes to choose and change ways to focus attention, learn, remember, and/or think. (capacity verb: *adopts*)	The learner's responses provide only a way of inferring that internal cognitive strategies were used. Among the cognitive strategies a learner may use are rehearsing (practicing), elaborating, organizing information, and metacognition. It is sometimes necessary to ask a learner to "think aloud" while performing a task in order to discover the cognitive processes the learner is using.	Given the task of learning a list of new Spanish vocabulary words, the learner is able to associate an English word with an "acoustical link" to help memorize the Spanish words' definitions.

Source: Table and excerpts adapted from *Principles of Instructional Design* (3rd ed., pp. 12, 57–68), by Robert M. Gagné, Leslie J. Briggs, & Walter W. Wager. Copyright © 1988. Reprinted by permission of Wadsworth, a division of Cengage.

FIGURE D.3 **Summary of the Quellmalz taxonomy.**

Classification	Definition	Illustration	Relation to Bloom taxonomy
Recall	Most tasks require that students recognize or remember key facts, definitions, concepts, rules, and principles. Recall questions require students to repeat verbatim or to paraphrase given information. To recall information, students need most often to rehearse or practice it, and then to associate it with other, related concepts. The Bloom taxonomy levels of knowledge and comprehension are subsumed here, since verbatim repetition and translation into the student's own words represent acceptable evidence of learning and understanding.	Who was the main character in the story?	Recall Comprehension
Analysis	In this operation, students divide a whole into component elements. Generally, the different part/whole relationships and the parts of cause/effect relationships that characterize knowledge within subject domains are essential components of more complex tasks. The components can be the distinctive characteristics of objects or ideas, or the basic actors of procedures or events. This definition of analysis is the same as that in the Bloom taxonomy.	What are the different story parts?	Analysis
Comparison	These tasks require students to recognize or explain similarities and differences. Simple comparisons require attention to one or a few very obvious attributes or component processes, while complex comparisons require identification of the differentiation among many attributes or component actions. This category relates to some of the skills in the Bloom level of analysis. The separate comparison category emphasizes the distinct information processing required when students go beyond breaking the whole into parts in order to compare similarities and differences. This is akin to the Bloom level of synthesis.	How was this story like the last one?	Analysis
Inference	Both deductive and inductive reasoning fall into this category. In deductive tasks, students are given a generalization and are required to recognize or explain the evidence that relates to it. Applications of rules and "if-then" relationships require inference. In inductive tasks, students are given the evidence or details and are required to come up with the generalization. Hypothesizing, predicting, concluding, and synthesizing all require students to relate and integrate information. Inductive and deductive reasoning relate to the Bloom levels of application and synthesis. Application of a rule is one kind of deductive reasoning; synthesis, putting parts together to form a generalization, occurs in both inductive and deductive reasoning.	What might be a good title for this story?	Application Synthesis
Evaluation	These tasks require students to judge quality, credibility, worth, or practicality. Generally, we expect students to use established criteria and explain how these criteria are or are not met. The criteria might be established rules of evidence, logic, or shared values. Bloom's levels of synthesis and evaluation are involved in this category. To evaluate, students must *assemble* and *explain* the interrelationship of evidence and reasons in support of their conclusion (synthesis). Explanation of criteria for reaching a conclusion is unique to evaluative reasoning.	Is this a good story? Why or why not?	Synthesis Evaluation

Source: Adapted from *Measuring Thinking Skills in the Classroom* (Table 1, pp. 8 and 19), revised edition, by R. J. Stiggins, E. Rubel, and E. Quellmalz. Copyright 1988. Washington, DC: National Educational Association. Adapted by permission of the NEA Professional Library.

FIGURE D.4 **Categories and subcategories of the Krathwohl et al. taxonomy of affective objectives with illustrative statements of objectives.**

Category	Definition	Learning Targets
1.0 Receiving (attending)		
1.1 Awareness	Be conscious of something . . . take into account a situation, phenomenon, object, or state of affairs. . . .	Develops awareness of aesthetic factors in dress, furnishings, architecture, city design, good art, and the like.
1.2 Willingness to receive	Being willing to tolerate a given stimulus, not to avoid it. . . . Willing to take notice of the phenomenon and give it . . . attention. . . .	Appreciation (tolerance) of cultural patterns exhibited by individuals from other groups—religious, social, economic, national, etc.
1.3 Controlled or selected attention	The control of attention, so that when certain stimuli are presented they will be attended to. . . . The favored stimulus is selected and attended to despite competing and detracting stimuli. . . .	Alertness toward human values and judgments on life as they are recorded in literature.
2.0 Responding		
2.1 Acquiescence in responding	"Obedience" or "compliance." . . . There is a passiveness so far as the initiation of behavior is concerned. . . .	Follows school rules on the playground.
2.2 Willingness to respond	The learner is sufficiently committed to exhibiting the behavior that he does so not just because of fear . . . but "on his own" or voluntarily. . . .	Volunteers to help classmates who are having difficulty with the science project.
2.3 Satisfaction in response	The behavior is accompanied by a feeling of satisfaction, an emotional response, generally of pleasure, zest, or enjoyment.	Finds pleasure in reading for recreation.
3.0 Valuing		
3.1 Acceptance of a value	The emotional acceptance of a proposition or doctrine on what one considers adequate ground. . . .	Continuing desire to develop the ability to speak and write effectively.
3.2 Preference for a value	The individual is sufficiently committed to a value to pursue it, to seek it out, to want it. . . .	Assumes responsibility for drawing reticent members of a group into conversation.
3.3 Commitment	"Conviction" and "certainty beyond a doubt." . . . Acts to further the thing valued, . . . to extend the possibility of . . . developing it, to deepen . . . involvement with it. . . .	Devotion to those ideas and ideals that are the foundation of democracy.
4.0 Organization		
4.1 Conceptualization of a value	The quality of abstraction or conceptualization is added (to the value or belief which permits seeing) . . . how the value relates to those he already holds or to new ones. . . .	Forms judgments as to the responsibility of society for conserving human and material resources.
4.2 Organization of a value system	To bring together a complex of values . . . into an ordered relationship with one another. . . .	Weighs alternative social policies and practices against the standards of the public welfare rather than the advantage of . . . narrow interest groups.
5.0 Characterization by a value or value complex		
5.1 Generalized set	Gives an internal consistency to the system of attitudes and values. . . . Enables the individual to reduce and order the complex world . . . and to act consistently and effectively in it.	Judges problems and issues in terms of situations, issues, purposes, and consequences involved rather than in terms of fixed, dogmatic precepts or emotional wishful thinking.
5.2 Characterization	One's view of the universe, one's philosophy of life, one's weltanschauung. . . .	Develops for regulation of one's personal and civic life a code of behavior based on ethical principles consistent with democratic ideals.

Source: Adapted from *Taxonomy of Educational Objectives: Book 2: Affective Domain* (pp. 176–185), by David R. Krathwohl, Benjamin S. Bloom, and Bertram B. Masia (Eds.). Published by Allyn and Bacon, Boston, MA. Copyright © 1964 by Pearson Education. Reprinted by permission of the publisher.

FIGURE D.5 **Categories and subcategories of the Harrow taxonomy of psychomotor and perceptual objectives.**

Classification Levels and Subcategories	Definitions	Learning Targets
1.00 Reflex Movements **1.10 Segmental Reflexes** **1.20 Intersegmental Reflexes** **1.30 Suprasegmental Reflexes**	Actions elicited without conscious volition in response to some stimuli.	Flexion, extension, stretch, postural adjustments.
2.00 Basic-Fundamental Movements **2.10 Locomotor Movements** **2.20 Non-Locomotor Movements** **2.30 Manipulative Movements**	Inherent movement patterns which are formed from a combining of reflex movements and are the basis for complex skilled movement.	Walking, running, jumping, sliding, hopping, rolling, climbing, pushing, pulling, swaying, swinging, stooping, stretching, bending, twisting, handling, manipulating, gripping, grasping finger movements.
3.00 Perceptual Abilities **3.10 Kinesthetic Discrimination** **3.20 Visual Discrimination** **3.30 Auditory Discrimination** **3.40 Tactile Discrimination** **3.50 Coordinated Abilities**	Interpretation of stimuli from various modalities providing data for the learner to make adjustments to his environment.	The *outcomes* of perceptual abilities are observable in *all purposeful* movement. Examples: Auditory—following verbal instructions. Coordinated—jumping rope, punting, catching.
4.00 Physical Abilities **4.10 Endurance** **4.20 Strength** **4.30 Flexibility** **4.40 Agility**	Functional characteristics of organic vigor which are essential to the development of highly skilled movement.	Distance running, distance swimming, weight lifting, wrestling, touching toes, back bend, ballet exercises, shuttle run, typing, dodgeball.
5.00 Skilled Movements **5.10 Simple Adaptive Skill** **5.20 Compound Adaptive Skill** **5.30 Complex Adaptive Skill**	A degree of efficiency when performing complex movement tasks which are based upon inherent movement patterns.	All skilled activities which build upon the inherent locomotor and manipulative movement patterns of classification level two.
6.00 Non-Discursive Communication **6.10 Expressive Movement** **6.20 Interpretive Movement**	Communication through bodily movements ranging from facial expressions through sophisticated choreographies.	Body postures, gestures, facial expressions, all efficiently executed skilled dance movements and choreographies.

Source: Adapted from *A Taxonomy of the Psychomotor Domain: A Guide for Developing Behavioral Objectives* (pp. 104–106), by A. J. Harrow, 1972, White Plains, NY: Longman. Reprinted by permission of the author.

FIGURE D.6A **The knowledge dimension of a revision of Bloom's *Taxonomy of Educational Objectives*.**

Major types and subtypes	Examples
A. Factual knowledge—The basic elements students must know to be acquainted with a discipline or solve problems in it	
AA. Knowledge of terminology	Technical vocabulary; musical symbols
AB. Knowledge of specific details and elements	Major national resources, reliable sources of information
B. Conceptual knowledge—The interrelationships among the basic elements within a larger structure that enable them to function together	
BA. Knowledge of classifications and categories	Periods of geological time; forms of business ownership
BB. Knowledge of principles and generalizations	Pythagorean theorem; law of supply and demand
BC. Knowledge of theories, models, and structures	Theory of evolution; structure of Congress
C. Procedural knowledge—How to do something; methods of inquiry; and criteria for using skills, algorithms, techniques, and methods	
CA. Knowledge of subject-specific skills and algorithms	Skills used in painting with watercolors; whole-number division algorithm
CB. Knowledge of subject-specific techniques and methods	Interviewing techniques; scientific method
CC. Knowledge of criteria for determining when to use appropriate procedures	Criteria used to determine when to apply a procedure involving Newton's second law; criteria used to judge the feasibility of using a particular method to estimate business costs

FIGURE D.6A (*continued*)

Major types and subtypes	Examples
D. Metacognitive knowledge—Knowledge of cognition in general as well as awareness and knowledge of one's own cognition	
Da. Strategic knowledge	Knowledge of outlining as a means of capturing the structure of a unit of subject matter in a textbook; knowledge of the use of heuristics
Db. Knowledge about cognitive tasks, including appropriate contextual and conditional knowledge	Knowledge of the types of tests particular teachers administer; knowledge of the cognitive demands of different tasks
Dc. Self-knowledge	Knowledge that critiquing essays is a personal strength, whereas writing essays is a personal weakness; awareness of one's own knowledge level business costs

Source: Adapted from *A Taxonomy for Learning, Teaching, and Assessing* (pp. 46, 67–68), by Lorin W. Anderson and David R. Krathwohl (Eds.). Published by Allyn and Bacon, Boston, MA. Copyright © 2001 by Pearson Education. Reprinted by permission of the publisher.

FIGURE D.6B **The cognitive process dimension of a revision of Bloom's *Taxonomy of Educational Objectives*.**

Categories & cognitive processes	Alternative names	Definitions and examples
1. Remember—Retrieve relevant knowledge from long-term memory		
1.1 Recognizing	Identifying	Locating knowledge in long-term memory that is consistent with presented material (e.g., recognize the dates of important events in U.S. history)
1.2 Recalling	Retrieving	Retrieving relevant knowledge from long-term memory (e.g., recall the dates of important events in U.S. history)
2. Understand—Construct meaning from instructional messages, including oral, written, and graphic communication		
2.1 Interpreting	Clarifying, paraphrasing, representing, translating	Changing from one form of representation (e.g., numerical) to another (e.g., verbal) (e.g., paraphrase important speeches and documents)
2.2 Exemplifying	Illustrating, instantiating	Finding a specific example or illustration of a concept or principle (e.g., give examples of various artistic painting styles)
2.3 Classifying	Categorizing, subsuming	Determining that something belongs to a category (e.g., concept or principle) (e.g., classify observed or described cases of mental disorders)
2.4 Summarizing	Abstracting, generalizing	Abstracting a general theme or major point(s) (e.g., write a short summary of the events portrayed on a videotape)
2.5 Inferring	Concluding, extrapolating, interpolating, predicting	Drawing a logical conclusion from presented information (e.g., when learning a foreign language, infer grammatical principles from examples)
2.6 Comparing	Contrasting, mapping, matching	Detecting correspondences between two ideas, objects, and the like (e.g., compare historical events to contemporary situations)
2.7 Explaining	Constructing models	Constructing a cause-and-effect model of a system (e.g., explain the causes of important 18th-century events in France)
3. Apply—Carry out or use a procedure in a given situation		
3.1 Executing	Carrying out	Applying a procedure to a familiar task (e.g., divide one whole number by another whole number, both with multiple digits)
3.2 Implementing	Using	Applying a procedure to an unfamiliar task (e.g., use Newton's second law in situations in which it is appropriate)
4. Analyze—Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or purpose		
4.1 Differentiating	Discriminating, distinguishing, focusing, selecting	Distinguishing relevant from irrelevant parts or important from unimportant parts of presented material (e.g., distinguish between relevant and irrelevant numbers in a mathematical word problem)
4.2 Organizing	Finding coherence, integrating, outlining, parsing, structuring	Determining how elements fit or function within a structure (e.g., structure evidence in a historical description into evidence for and against a particular historical explanation)

FIGURE D.6B (*continued*)

Categories & cognitive processes	Alternative names	Definitions and examples
4.3 Attributing	Deconstructing	Determine a point of view, bias, values, or intent underlying presented material (e.g., determine the point of view of the author of an essay in terms of his or her political perspective)
5. Evaluate—Make judgments based on criteria and standards		
5.1 Checking	Coordinating, detecting, monitoring, testing	Detecting inconsistencies or fallacies within a process or product; determining whether a process or product has internal consistency; detecting the effectiveness of a procedure as it is being implemented (e.g., determine if a scientist's conclusions follow from observed data)
5.2 Critiquing	Judging	Detecting inconsistencies between a product and external criteria, determining whether a product has external consistency; detecting the appropriateness of a procedure for a given problem (e.g., judge which of two methods is the best way to solve a given problem)
6. Create—Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure		
6.1 Generating	Hypothesizing	Coming up with alternative hypotheses based on criteria (e.g., generate hypotheses to account for an observed phenomenon)
6.2 Planning	Designing	Devising a procedure for accomplishing some task (e.g., plan a research paper on a given historical topic)
6.3 Producing	Constructing	Inventing a product (e.g., build habitats for a specific purpose)

Source: Adapted from *A Taxonomy for Learning, Teaching, and Assessing* (pp. 46, 67–68), by Lorin W. Anderson and David R. Krathwohl (Eds.). Published by Allyn and Bacon, Boston, MA. Copyright © 2001 by Pearson Education. Reprinted by permission of the publisher.

FIGURE D.7 **Categories of learning targets derived from the dimensions of learning model.**

Declarative Knowledge

Procedural Knowledge

Complex Thinking

A. Effectively translates issues and situations into meaningful tasks that have a clear purpose.

B. Effectively uses a variety of complex reasoning strategies.

REASONING STRATEGY 1: COMPARISON Comparison involves describing the similarities and differences between two or more items. The process includes three components that can be assessed:

a. Selects appropriate items to compare.

b. Selects appropriate characteristics on which to base the comparison.

c. Accurately identifies the similarities and differences among the items, using the identified characteristics.

REASONING STRATEGY 2: CLASSIFICATION Classification involves organizing items into categories based on specific characteristics. The process includes four components that can be assessed:

a. Selects significant items to classify.

b. Specifies useful categories for the items.

c. Specifies accurate and comprehensive rules for category membership.

d. Accurately sorts the identified items into the categories.

REASONING STRATEGY 3: INDUCTION Induction involves creating a generalization from implicit or explicit information and then describing the reasoning behind the generalization. The process includes three components that can be assessed:

a. Identifies elements (specific pieces of information or observations) from which to make inductions.

b. Interprets the information from which inductions are made.

c. Makes and articulates accurate conclusions (inductions) from the selected information or observations.

FIGURE D.7 (*continued*)

REASONING STRATEGY 4: DEDUCTION Deduction involves identifying implicit or explicit generalizations or principles (premises) and then describing their logical consequences. The process includes three components that can be assessed:

a. Identifies and articulates a deduction based on important and useful generalizations or principles implicit or explicit in the information.

b. Accurately interprets the generalizations or principles.

c. Identifies and articulates logical consequences implied by the identified generalizations or principles.

REASONING STRATEGY 5: ERROR ANALYSIS Error analysis involves identifying and describing specific types of errors in information or processes. It includes three components that can be assessed:

a. Identifies and articulates significant errors in information or in process.

b. Accurately describes the effects of the errors on the information or process.

c. Accurately describes how to correct the errors.

REASONING STRATEGY 6:CONSTRUCTING SUPPORT Constructing support involves developing a well-articulated argument for or against a claim. The process includes three components that can be assessed:

a. Accurately identifies a claim that requires support rather than a fact that does not require support.

b. Provides sufficient or appropriate evidence for the claim.

c. Adequately qualifies or restricts the claim.

REASONING STRATEGY 7: ABSTRACTING Abstracting involves identifying and explaining how the abstract pattern in one situation or set of information is similar to or different from the abstract pattern in another situation or set of information. The process includes three components that can be assessed:

a. Identifies a significant situation or meaningful information that is a useful subject for the abstracting process.

b. Identifies a representative general or abstract pattern for the situation or information.

c. Accurately articulates the relationship between the general or abstract pattern and another situation or set of information.

REASONING STRATEGY 8: ANALYZING PERSPECTIVES Analyzing perspectives involves considering one perspective on an issue and the reasoning behind it as well as an opposing perspective and the reasoning behind it. The process includes three components that can be assessed:

a. Identifies an issue on which there is disagreement.

b. Identifies one position on the issue and the reasoning behind it.

c. Identifies an opposing position and the reasoning behind it.

REASONING STRATEGY 9: DECISION MAKING Decision making involves selecting among apparently equal alternatives. It includes four components that can be assessed:

a. Identifies important and appropriate alternatives to be considered.

b. Identifies important and appropriate criteria for assessing the alternatives.

c. Accurately identifies the extent to which each alternative possesses each criteria.

d. Makes a selection that adequately meets the decision criteria and answers the initial decision question.

REASONING STRATEGY 10: INVESTIGATION Investigation is a process involving close examination and systematic inquiry. There are several basic types of investigation:

- *Definitional investigation*: Constructing a definition or detailed description concept for which such a definition or description is not readily available or accepted.
- *Historical investigation*: Constructing an explanation for some past event for which an explanation is not readily available or accepted.
- *Projective investigation*: Constructing a scenario for some future event or hypothetical past event for which a scenario is not readily available or accepted.

Each type of investigation includes three components that can be assessed:

a. Accurately identifies what is already known or agreed on about the concept (definitional investigation), the past event (historical investigation), or the future event (projective investigation).

b. Identifies and explains the confusions, uncertainties, or contradictions about the concept (definitional investigation), the past event (historical investigation), or the future event (projective investigation).

c. Develops and defends a logical and plausible resolution to the confusions, uncertainties, or contradictions about the concept (definitional investigation), the past event (historical investigation), or the future event (projective investigation).

FIGURE D.7 (*continued*)

REASONING STRATEGY 11: PROBLEM SOLVING Problem solving involves developing and testing a method or product for overcoming obstacles or constraints to reach a desired outcome. It includes four components that can be assessed:

a. Accurately identifies constraints or obstacles.
b. Identifies viable and important alternatives for overcoming the constraints or obstacles.
c. Selects and adequately tries out alternatives.
d. If other alternatives were tried, accurately articulates and supports the reasoning behind the order of their selection, and the extent to which each overcame the obstacles or constraints.

REASONING STRATEGY 12: EXPERIMENTAL INQUIRY Experimental inquiry involves testing hypotheses that have been generated to explain phenomenon. It includes four components that can be assessed:

a. Accurately explains the phenomenon initially observed using appropriate and accepted facts, concepts, or principles.
b. Makes a logical prediction based on the facts, concepts, or principles underlying the explanation.
c. Sets up and carries out an activity or experiment that effectively tests the prediction.
d. Effectively evaluates the outcome of the activity or experiment in terms of the original explanation.

REASONING STRATEGY 13: INVENTION Invention involves developing something unique or making unique improvements to a process to satisfy an unmet need. It includes four components that can be assessed:

a. Identifies a process or product to develop or improve to satisfy an unmet need.
b. Identifies rigorous and important standards or criteria the invention will meet.
c. Makes detailed and important revisions in the initial process or product.
d. Continually revises and polishes the process or product until it reaches a level of completeness consistent with the criteria or standards identified earlier.

Information Processing

A. Effectively interprets and synthesizes information.
B. Effectively uses a variety of information-gathering techniques and resources.
C. Accurately assesses the value of information.
D. Recognizes where and how projects would benefit from additional information.

Effective Communication

A. Expresses ideas clearly.
B. Effectively communicates with diverse audiences.
C. Effectively communicates in a variety of ways.
D. Effectively communicates for a variety of purposes.
E. Creates quality products.

Collaboration/Cooperation

A. Works toward the achievement of group goals.
B. Demonstrates effective interpersonal skills.
C. Contributes to group maintenance.
D. Effectively performs a variety of roles within a group.

Habits of Mind

A. Is aware of own thinking.
B. Makes effective plans.
C. Is aware of and uses necessary resources.
D. Evaluates the effectiveness of own actions.
E. Is sensitive to feedback.
F. Is accurate and seeks accuracy.
G. Is clear and seeks clarity.

FIGURE D.7 ***(continued)***

H. Is open-minded.
I. Restrains impulsivity.
J. Takes a position when the situation warrants it.
K. Is sensitive to the feelings and level of knowledge of others.
L. Engages intensively in tasks even when answers or solutions are not immediately apparent.
M. Pushes the limits of own knowledge and ability.
N. Generates, trusts, and maintains own standards of evaluation.
O. Generates new ways of viewing a situation outside the boundaries of standard convention.

Source: Adapted from *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model* (pp. 65–93), by R. J. Marzano, D. Pickering, and J. McTighe, 1993. Alexandria, VA: Association for Supervision and Curriculum Development. (Copyright by McREL, 4601 DTC Boulevard #500, Denver, CO 80237.) Adapted by permission.

Implementing the Principles of Universal Design via Technology-Based Testing

FIGURE E.1

What are the principles of universal design?	What does the principle mean for testing?	How can the principle be implemented?
Principle 1: Equitable Use	Testing materials, strategies, and environments are designed so that they are useful, appealing, and safe for *all* to use. They are respectful of individual differences and are used by *all* learners in similar or equivalent ways and In different contexts.	• Use the principles of typographic and visual design • Pair text with culturally and age appropriate visuals • Provide instructional feedback to students
Principle 2: Flexible Use	Testing materials, strategies, and environments are designed so that they accommodate individual preferences and abilities. They are flexible In terms of providing choices of the methods and pace of use.	• Offer options to students about the technology they use to take tests • Give students choices about pace, location, and sequence of the test administration • Allow students to take tests/quizzes multiple times
Principle 3: Simple and Intuitive	Testing materials, strategies, and environments are designed so that they are easy for *all* to use and understand. Their use is not dependent on one's experience, prior knowledge, language and literacy skills, and other learning preferences and abilities.	• Use software/Web sites to check and enhance the readability of tests • Provide second language learners with access to bilingual resources • Embed varied visual supports that are current, age appropriate, and culturally sensitive
Principle 4: Perceptible Information	Testing materials, strategies, and environments are designed so that they communicate essential information to *all* using multiple formats, backgrounds with sufficient contrasts, legible text guidelines, compatible teaching and testing techniques, and assistive technology devices.	• Use the principles of typographic and visual design to prepare legible and readable testing materials • Use technology to present test directions and items (e.g., screen/text-reading programs) • Provide visual supports • Offer prompts and cues to help students understand test directions and items
Principle 5: Tolerance for Error	Testing materials, strategies, and environments are designed to minimize errors, adverse consequences, and unintentional actions. They provide safeguards and warnings to assist *all* in using them safely and efficiently.	• Provide learning strategy access and reminders • Embed feedback and error minimization techniques into tests • Allow students to use word processors, spellcheckers, word cueing and prediction, dictionaries and thesauri, and grammar checkers • Teach technology-based test-taking skills

From Appendix E of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FIGURE E.1 (*continued*)

What are the principles of universal design?	What does the principle mean for testing?	How can the principle be Implemented?
Principle 6: Low Physical Effort	Testing materials, strategies, and environments are designed to be used comfortably and without much physical effort by *all*. They allow all to use them with a range of reasonable physical actions and do not require repetitive actions or sustained physical effort.	• Provide students with the technology they need to take tests (e.g., voice-activation, augmentative communication, and low-tech devices, etc.)
Principle 7: Size and Space Approach and Use	Testing materials, strategies, and environments are designed for use by *all* regardless of one's body size, posture, and mobility. They allow all users to see, reach, and activate important features and information and offer sufficient space for assistive technology devices and personal assistance.	• Provide students with ergonomic and alternative keyboards, an adapted mouse, keyguards, on-screen keyboarding, visual and auditory warnings, and highlighted mouse visibility and movement • Format tests appropriately
Principle 8: Community of Learners	Testing materials, strategies, and environments promote socialization and communication.	• Present tests /quizzes using technology-based and collaborative game formats
Principle 9: Inclusive Environment	Testing materials, strategies, and environments foster acceptance and belonging.	• Use branching to tailor tests to students' skill levels • Motivate students by providing choices regarding the frequency and type of feedback they receive

Source: From "Using Technology to Create and Administer Accessible Tests," by S. Salend, 2009, *TEACHING Exceptional Children, 41*(3), p. 42. Copyright © 2009 by the Council for Exceptional Children, Inc. http://www.cec.sped.org. All rights reserved. Reproduced with permission.

Assessment of Metacognition

DEFINITION OF METACOGNITION

There are many facets to teaching students to use thinking skills. In Chapter 2 we discussed several frameworks for identifying thinking skills that should be incorporated into your teaching and assessment practices. One broad area of thinking that has received considerable attention in recent years from researchers and curriculum specialists is students' abilities to monitor and control their own thinking in relation to the cognitive tasks they are performing. Monitoring and controlling one's own thinking processes are complex skills themselves. The cluster of such related skills is known as metacognitive skills.

Metacognition is defined as "one's knowledge concerning one's own cognitive processes and products or anything related to them. . . . For example, I am engaging in metacognition . . . if I notice that I am having more trouble learning A than B; if it strikes me that I should double check C before accepting it as a fact. . . . Metacognition refers, among other things, to the active monitoring and consequent regulation and orchestration of these processes . . . usually in the service of some concrete goal or objective" (Flavell, 1976, p. 232).

As can be surmised, students engage in metacognitive thinking when they are aware of their thoughts as they perform specific learning activities and then use this awareness to control what they are doing (Marzano et al., 1988).

TYPES OF METACOGNITIVE SKILLS

The metacognitive cluster of skills can be organized in several ways. The Marzano et al. (1988) organization gives a brief overview of this domain of learning targets.

I. *Self-regulation skills* are used by students when they are aware that they can control their commitment, attitudes, and attention toward academic tasks.

   A. *Commitment to an academic task* is a student's conscious decision to choose to do the task, whether or not it is fun for the student.

   B. *Positive attitude toward an academic task* is a student's belief that she can perform the task and that the main determiner of success on it is her own efforts, not luck, natural talent, or help from others.

   C. *Controlling attention to the requirements of an academic task* occurs when a student recognizes that he must control the level and focus of his attention to match the requirements of the task to be performed.

II. *Types of knowledge* used by students must be appropriate for performing the academic task at hand.

   A. *Declarative knowledge* is exhibited when a student knows what needs to be done, knows factual information, or knows that something is to be done.

   B. *Procedural knowledge* is exhibited when a student is able to perform a task or to apply strategies to complete tasks.

   C. *Conditional knowledge* is exhibited when a student is aware of why certain procedures or strategies are used or in what circumstances one procedure or strategy is preferred over another.

III. *Executive control skills* are used by students when they evaluate, plan, and check their own progress in completing an academic task.

   A. *Evaluation skills* are used when a student assesses her current state of knowledge before, during, and at the completion of an academic task; identifies available and still-needed resources for

From Appendix F of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

completing the academic task; and identifies the goals and subgoals of the academic task.

B. *Planning skills* are used by a student before and during the completion of an academic task when the student deliberately chooses procedures and strategies to do the task.

C. *Regulating processes skills* are used by a student while completing an academic task when the student monitors his progress toward completing the task successfully.

These categories of skills are not hierarchical and, in practice, students usually use them in combination to complete academic tasks.

ASSESSING METACOGNITION WITH PAPER-AND-PENCIL INSTRUMENTS

Suggestions for how to model and teach these skills are found in other sources (Good & Brophy, 2002; Marzano et al., 1988). Here we give some examples of simple ways to assess students' perceptions of whether they use these skills.

Teachers have found this type of assessment information useful for planning instruction (Tittle, 1989; Tittle, Hecht, & Moore, 1993). To create an instrument, you need to identify a specific instructional activity on which to focus the items. For example, you may wish to focus on students' metacognitions during class, while working with others, while doing homework or other assignments and projects, or when they complete tests or other assessment activities used for summative evaluation (Tittle et al., 1993). Then, using the subcategories of metacognitive skills, write statements describing a student's thoughts, beliefs, or awareness about the specific type of activity. Write both positive and negative statements (i.e., "good" and "poor" metacognitions) for each category. Give each student a copy of the list and ask him or her to indicate how often he or she does the things in each statement.

Figure F.1 shows some examples of statements related to various metacognitions that may occur when students

FIGURE F.1 Examples of positively and negatively phrased items that assess how students report using metacognitions when preparing a social studies research paper.

Directions to students: These questions ask about how often you do some things when you write a research paper in social studies. Circle the number that tells how often you do each thing.

Never or almost never	Sometimes	Often	Always or almost always	Don't Know
1	2	3	4	DK

1.	When the paper is not a lot of fun, I work very hard to do a good job on it. [I.A, +]	1	2	3	4	D K
2.	When the paper is not a lot of fun, I do not work very hard on it. [I.A, −]	1	2	3	4	D K
3.	I do a good job on the paper even though I am less talented than other students. [I.B, +]	1	2	3	4	D K
4.	I have to get really lucky to do a good job on the paper. [I.B, −]	1	2	3	4	D K
5.	When I read articles related to my paper, I read only those parts related to my topic. [I.C, +]	1	2	3	4	D K
6.	When reading articles about the topic of my paper, I give equal attention to everything in the article. [I.C, −]	1	2	3	4	D K
7.	My research papers have an introductory section that tells why the topic is important. [II.A, +]	1	2	3	4	D K
8.	My research papers list the facts about the topic but do not give my interpretation of the meaning of the facts. [II.A, −]	1	2	3	4	D K
9.	I make tables in my research papers to compare information on the topic. [II.B, +]	1	2	3	4	D K
10.	I do not use note cards when preparing my research paper. [II.B, −]	1	2	3	4	D K
11.	Before I decide to use a graph or a chart I ask myself which idea in the paper it supports. [II.C, +]	1	2	3	4	D K
12.	I use lots of graphs or charts in my research reports. [II.C, −]	1	2	3	4	D K
13.	One of the first things I do when I start my paper assignment is to make a list of what I already know about the topic. [III.A, +]	1	2	3	4	D K
14.	Before I do anything else on the paper I go to the library to find all the books and articles about my topic. [III.B, −]	1	2	3	4	D K
15.	After I complete my research paper I ask myself what I learned about the topic. [III.A, +]	1	2	3	4	D K
16.	After I complete my research paper, I do not think about the topic anymore. [III.A, −]	1	2	3	4	D K
17.	When I am ready to begin collecting information, I ask myself what sources would be best to use first. [III.B, +]	1	2	3	4	D K
18.	No matter what the topic of my paper, I go first to the encyclopedia to look up the topic. [III.B, −]	1	2	3	4	D K
19.	While I am writing the paper I think about whether it meets the criteria for a good research report. [III.C, +]	1	2	3	4	D K
20.	As soon as I have a little information on the topic, I begin writing the paper. [III.C, −]	1	2	3	4	D K

Notes: Codes in brackets refer to outline in the text.

work on a social studies research paper. Students are asked to identify how often they engage in the thoughts, actions, or beliefs listed. The statements are arranged to follow the outline of metacognitive skills discussed earlier. The statements are in pairs: The odd-numbered member of the pair is a positive statement, and the even-numbered one is a negative statement. The codes in the brackets identify the skill in the outline and whether the statement is positively or negatively worded. Remember that Figure F.1 is just a list of examples, not a sample instrument per se. In an actual instrument you would scramble the order of positive and negative statements so as not to have a pattern, omit the codes, and have more than two statements per category. Also, you might not assess some categories because of the nature of the particular activity on which you are focusing. Note that such an instrument may be inappropriate for primary children, whose reading skills may not be sufficient to understand it.

Examples of Alternative Blueprints for a Summative Unit Assessment

FIGURE G.1 **A checklist for judging the quality of a teacher's plan for a summative unit assessment.**

1.	Does your plan clarify the purpose(s) of the assessment and what you expect it to tell you about each student?	Yes	No
2.	Does your plan indicate the main subject-matter topics and performances you want to assess?	Yes	No
3.	Will your plan help you to judge whether the assessment tasks match the major content topics and learning targets you have specified?	Yes	No
4.	Have you clearly identified the elements of knowledge and performance that *all* students need to know?	Yes	No
5.	Does your plan give the most important learning targets the heaviest weights in the total score? Are the least important learning targets given the least weight? (You may wish to give certain tasks more weight than others.)	Yes	No
6.	Do you know what kind(s) of assessment tasks should be used to assess each content-thinking skill combination? Are these tasks the best ways to assess the combination?	Yes	No
7.	Have you estimated the amount of time students need to complete this assessment? Is this estimated time realistic?	Yes	No
8.	Have you estimated the amount of time you will need to evaluate the students' responses? (Consider how this time might be shortened, without reducing the validity of the results, by changing some of the tasks, rearranging tasks on a page, or using the capabilities of a microcomputer or other scoring device.)	Yes	No

Note: Revise your assessment plan if you answered no to one or more of the questions in the checklist.

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 26), by A. J. Nitko and T. C. Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of the authors.

From Appendix G of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FIGURE G.2 Complete specifications with modified taxonomy headings.

Content outline	Recalling information taught or read	Applying knowledge in situations very similar to those taught	Applying knowledge in a new or novel context
I. *Basic Parts of Cell* A. *Nucleus* B. *Cytoplasm* C. *Cell membrane*	1. *Names and tells functions of each part of cell*	8. *Labels parts of cell shown on a line drawing*	11. *Given photographs of actual plant and animal cells, labels the parts*
40% of Total = 8 pts	37% of Row = 3 pts	37% of Row = 3 pts	26% of Row = 2 pts
II. *Plant vs. Animal Cells* A. *Similarities* B. *Differences* 1. *cell wall vs. membrane* 2. *food manufacture*	2. *Explains differences between plant and animal cells* 3. *Describes the cell wall and cell membrane*		
10% of Total = 2 pts	100% of Row = 2 pts	___% of Row = __ pts	___% of Row = __ pts
III. *Cell Membrane* A. *Living nature of* B. *Diffusion* C. *Substances diffused by cells*	4. *Lists substances diffused and not diffused by cell membranes* 5. *Gives definition of diffusion*	9. *Distinguishes between diffusion and oxidation*	
20% of Total = 4 pts	75% of Row = 3 pts	25% of Row = 1 pts	___% of Row = __ pts
IV. *Division of Cells* A. *Phases in division* B. *Chromosomes and DNA* C. *Plant vs. animal cell division*	6. *Gives definitions of division, chromosomes, and DNA* 7. *States differences between plant and animal cell division*	10. *Given the numbers of chromosomes in a cell before division, states the number in each cell after division*	
30% of Total = 6 pts	67% of Row = 4 pts	33% of Row = 2 pts	___% of Row = __ pts

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 4) by A. J. Nitko and T.-C. Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of the authors.

FIGURE G.3 Blueprint without objectives stated.

Content outline	Recalling information taught or read	Applying knowledge in situations very similar to those taught	Applying knowledge in a new or novel context
I. *Basic Parts of Cell* A. *Nucleus* B. *Cytoplasm* C. *Cell membrane*	*1 item, scored 0–3 (short-answer)*	*1 item, scored 0–3 (label parts) of cell drawing*	*2 items, each scored 0–1 (label parts of cell photographs)*
40% of Total = 8 pts	37% of Row = 3 pts	37% of Row = 3 pts	26% of Row = 2 pts
II. *Plant vs. Animal Cells* A. *Similarities* B. *Differences* 1. *cell wall vs. membrane* 2. *food manufacture*	*2 items, each scored 0–1 (short-answer)*		
10% of Total = 2 pts	100% of Row = 2 pts	___% of Row = __ pts	___% of Row = __ pts
III. *Cell Membrane* A. *Living nature of* B. *Diffusion* C. *Substances diffused by cells*	*2 items, one scored 0–2, the other scored 0–1 (short-answer)*	*2 items, each scored 0–1 (multiple-choice)*	
20% of Total = 4 pts	75% of Row = 3 pts	25% of Row = 1 pts	___% of Row = __ pts
IV. *Division of Cells* A. *Phases in division* B. *Chromosomes and DNA* C. *Plant vs. animal cell division*	*4 items, each scored 0–1 (definitions, short-answer)*	*1 item, scored 0–1 (short-answer)*	
30% of Total = 6 pts	67% of Row = 4 pts	33% of Row = 2 pts	___% of Row = __ pts

Source: Adapted from *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 4) by A. J. Nitko and T.-C. Hsu, 1987, Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Adapted by permission of the authors.

FIGURE G.4 **Blueprint using only a list of learning targets.**

Objectives of the unit	Number of items	Number of points
1. Names and tells functions of each cell part.	1	3
2. Explains differences between plant and animal cells.	1	1
3. Describes the cell wall and cell membrane.	1	1
4. Lists substances diffused and not diffused through cell membrane.	1	2
5. Gives definition of diffusion.	1	1
6. Gives definition of division, chromosomes, and DNA.	3	3
7. States differences between plant and animal cell division.	1	1
8. Labels parts of a cell when shown a line drawing.	3	3
9. Distinguishes between diffusion, oxidation, and fission.	2	2
10. Given the number of chromosomes in a cell before division, states the number in each cell after division.	1	1
11. Given photographs of plant and animal cells, identifies parts of cells without using prompts.	2	2
	Total points = 17	Total points = 20

Source: From *Teacher's Guide to Better Classroom Testing: A Judgmental Approach* (p. 39), by A. J. Nitko and T.-C. Hsu, 1987. Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Reprinted by permission.

FIGURE G.5 **Blueprint using only a content listing.**

Major topics of unit	Number of items	Number of points per item	Total number of points
I. Basic Parts of a Cell			
A. Nucleus	2	2	4
B. Cytoplasm	3	1	3
C. Cell membrane	1	1	1
	subtotal = 6		subtotal = 8
II. Plant vs. Animal Cells			
A. Similarities	1	1	1
B. Differences			
1. cell wall vs. cell membrane	1	1	1
2. food manufacture	0		0
	subtotal = 2		subtotal = 2
III. Cell Membrane			
A. Living nature of	0		0
B. Diffusion	2	1	2
C. Examples of different substances	1	2	2
	subtotal = 3		subtotal = 4
IV. Division of Cells			
A. Phases in division	2	1	2
B. Role of chromosomes and DNA	2	1	2
C. Plant vs. animal cell division	2	1	2
	subtotal = 6		subtotal = 6
	Total test items = 17		Total test points = 20

Source: From *Teacher's Guide to Better Classroom Testing, A Judgmental Approach* (p. 38), by A. J. Nitko and T.-C. Hsu, 1987. Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh. Reprinted by permission of the authors.

Scoring Guide for Oregon's Writing Assessment

WRITING SCORING GUIDE: MIDDLE SCHOOL STUDENT VERSION

Ideas and Content **Communicating knowledge of the topic, including relevant examples, facts, anecdotes and details**		
6 **The writing is exceptionally clear, focused and interesting. It holds the reader's attention. Main ideas stand out and are developed by strong support and rich details that fit the audience and purpose. The writing has** • a clear focus and control. • main idea(s) that stand out. • details that are on topic and carefully selected; when needed, use of resources provides strong, accurate, believable support. • an appropriate amount of detail (not too much or too little) to support an in-depth explanation or exploration of the topic; the writing makes connections and shares insights. • main ideas and selected details that fit the purpose and hold the reader's attention from beginning to end.	5 **The writing is clear, focused and interesting. It holds the reader's attention. Main ideas stand out and are developed by supporting details that fit the audience and purpose. The writing has** • a clear focus and control. • main idea(s) that stand out. • details that are on topic and carefully selected; when needed, use of resources provides strong, accurate, believable support. • an appropriate amount of detail (not too much or too little) to support a thorough explanation or exploration of the topic; the writing makes connections and shares insights. • main ideas and selected details that fit the purpose and hold the reader's attention from beginning to end.	4 **The writing is clear and focused. The reader can easily understand the main ideas. Support is present, but may be limited or somewhat general. The writing has** • a clear purpose. • clear main ideas. • details that are on topic, but may be too general or limited; when needed, resources are used to provide accurate support. • details that may sometimes be too many or too few for a thorough explanation or exploration of the topic; some connections and insights may be present. • main ideas and selected details that fit the purpose and hold the reader's attention most of the time from beginning to end.
3 **The writing has main idea(s), but they may be too broad or simplistic. Supporting detail is often too limited, overly general, or sometimes off the topic. The writing has** • a purpose that is easy to find. • main idea(s) that are easy to find but overly obvious or predictable; main points or conclusions repeat ideas often heard. • support of main ideas, but there aren't enough supporting details, or they are too general, predictable, or somewhat off topic. • details that may not be based on reliable resources; may be based on clichés, stereotypes, or sources of information that are biased, uninformed, or unreliable.	2 **The writing has main idea(s), but they are undeveloped, and the purpose is somewhat unclear. The writing has** • an unclear purpose that requires the reader to guess the main ideas. • minimal development, lacking details. • details, when included, are not well connected to the main ideas and clutter the paper. • details that are frequently repeated.	1 **The writing lacks main idea(s) or purpose. The writing has** • ideas that are very limited or simply unclear. • few or no attempts to develop ideas; the paper is too short to demonstrate the development of an idea.

From Appendix H of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

Scoring Guide for Oregon’s Writing Assessment

Voice **Expressing ideas in an engaging and credible way for audience and purpose**		
6 **The writer has chosen an appropriate voice for the topic,purpose and audience and shows a deep sense of involvement with the topic. The writing is interesting and sincere. The writing has** • an effective level of closeness to the audience or distance from it (e.g., a narrative should have a strong personal voice, while a research paper may require a more objective voice; both should be lively or interesting). • an exceptionally strong sense of purpose and audience. • a sense that the topic has come to life; when appropriate, shows use of originality, liveliness, honesty, conviction, excitement, humor, suspense and/or use of outside resources.	**5** **The writer has chosen an appropriate voice for the topic, purpose and audience and shows involvement with the topic. The writing is interesting and seems sincere. The writing has** • an appropriate level of closeness to the audience or distance from (e.g., a narrative should have a strong personal voice, while a research report may require a more objective voice; both could be lively or interesting.) • a strong sense of purpose and audience. • a sense that the topic has come to life; when appropriate, the writing shows originality, liveliness, honesty, conviction, excitement, humor, suspense and/or use of outside resources.	**4** **A voice is present,and there is a sense of involvement with the topic. In places, the writing is interesting and seems sincere. The writing has** • a questionable or inconsistent level of closeness or distance from the audience. • a sense of purpose and audience but may not use a consistently appropriate voice. • originality, liveliness, humor and/or use of outside resources, when appropriate; however, at times voice may be too casual or formal.
3 **The writer doesn’t seem particularly involved with the topic or may seem either too personal or too impersonal. The writing has** • a voice that doesn’t seem to match the topic, purpose, and audience. • a limited sense that the paper was written for a particular audience. • a sense in places of the writer behind the words; however, this may shift or disappear a line or two later. • limited ability to shift from a casual, informal voice to one that is more objective when that is necessary.	**2** **The writing provides little sense of involvement or evidence of a suitable voice. The writing has** • little or no sense that the writer cares about the topic; the writing is largely flat, lifeless, stiff, or mechanical. • little or no awareness of matching the topic, purpose and audience. • little or no sense of the writer behind the words; there are only a few places where the reader and writer can feel a connection. • a voice that is likely to be overly formal or overly personal.	**1** **The writing lacks a sense of involvement and a suitable voice. The writing has** • no sense that the writer cares about the topic; the writing is flat, lifeless, stiff, or mechanical. • no sense that the piece was written for an audience. • no hint of the writer behind the words; there are few if any places where the reader feels connected to the writer. The writing doesn’t get the reader involved.

Scoring Guide for Oregon’s Writing Assessment

Organization **Structuring information in logical sequence, making connections and transitions among ideas, sentences, and paragraphs**		
6 **The organization makes the central idea(s) and supporting details clear. The order and structure are strong and move the reader easily through the writing. The writing has** • effective (and sometimes creative) ideas, details, and examples in an order that is easy to follow. • a strong and inviting introduction that draws the reader in and a strong conclusion that leaves the reader satisfied • smooth, effective transitions that tie together ideas, sentences and paragraphs; the reader can move easily from one part to the next. • details placed where they work well and make the most sense.	**5** **The organization helps clarify the central idea(s) and supporting details. The order and structure are strong and move the reader through the writing. The writing has** • ideas, details, and examples in an order that makes sense and is easy to follow. • an inviting introduction that draws the reader in and a conclusion that leaves the reader satisfied. • smooth transitions that tie together ideas, sentences, and paragraphs; the reader can move easily from one part to the next. • details placed where they work well and make the most sense.	**4** **The organization is clear and functional. Order and structure are present, but may seem like a formula. The writing has** • clear sequencing. • an organization that may be predictable. • an introduction that is recognizable but may not be especially inviting;a developed conclusion that is functional but may seem repetitive and ordinary. • transitions that work but they may be awkward or common. • a body that is easy to follow with details that fit where placed. • an organization which helps the reader, despite some weaknesses.
3 **An attempt to organize the writing has been made, but it doesn’t work well in places or is too obvious. The writing has** • attempts to put ideas in order, but the order is sometimes unclear. • a beginning and an ending, but they are either too short or too obvious (e.g., “My topic is ...”; “These are all the reasons that ...”) • a limited number of transitional words that are used too many times (e.g., “and,” “then,” “ but,” “ so,”“or,”“for,” “ yet,” numbering) • a structure that is too obvious, almost like a formula. • details that seem out of order and confuse the reader. • an organization that helps the reader in some places but breaks down in others.	**2** **The writing lacks a clear organizational structure. An occasional attempt at organizing is made, but the writing is difficult to follow and the reader has to reread large sections. The writing may seem incomplete. The writing has** • some attempts to organize ideas, but the order does not make the meaning clear. • a missing or extremely undeveloped introduction, body, or conclusion. • few or no transitions; when present they are ineffective or overused. • details are randomly placed; the reader is frequently confused.	**1** **The writing doesn’t hold together; the writing seems haphazard and disjointed. Even after rereading, the reader is still confused. The writing has** • ideas that are not in a clear or logical order. • no recognizable beginning or ending. • few or no transitions. • arrangement and pace of ideas that either drag or feel rushed.

Scoring Guide for Oregon's Writing Assessment

Sentence Fluency **Developing flow and** **rhythm of sentences**		
6 **The writing has an effective flow that is smooth and natural. The sentences are put together so they are consistently varied and interesting. The sentences make the piece easy and interesting to read. The writing has** • a natural, fluent sound; it glides along with one sentence flowing effortlessly into the next. • extensive variation in sentence lengths, patterns, and beginnings that make the writing interesting. • a sentence structure that helps the reader understand the text by highlighting key ideas and relationships. • strong control over sentence structure; if fragments are used at all, they work well. • natural-sounding dialogue, if dialogue is used at all.	**5** **The writing has a smooth, natural flow. Sentences are put together so they are varied and interesting. The sentences make the piece easy and interesting to read aloud. The writing has** • a natural, fluent sound; it glides along with one sentence flowing into the next. • a variety of sentence lengths, patterns, and beginnings that make the writing interesting. • sentence structure that helps the reader understand the meaning. • control over sentence structure; if fragments are used at all, they work well. • natural-sounding dialogue, if dialogue is used at all.	**4** **The writing flows; however, connections between phrases or sentences may be less than fluid. Sentences are somewhat varied, making oral reading easy. The writing has** • a natural sound; the reader can move easily through the piece, although it may lack a sense of rhythm. • some repeated sentence lengths, patterns and beginnings that detract somewhat from overall impact. • strong control over simple sentences; less control over more complex sentences. If fragments are used at all, they are usually effective. • dialogue, if used at all, that usually sounds natural but can sound artificial.
3 **The writing tends to be choppy rather than smooth. Sometimes awkward constructions force the reader to slow down or reread. The writing has** • some passages that are easy to read aloud and some that are choppy. • some variety in sentence lengths, patterns, and beginnings, although a few are used repeatedly. • simple sentence used correctly, but more complex sentences may have problems; if fragments are used, they may not be effective. • sentences that are correct, but are not very interesting or appealing. • dialogue that may sound unnatural or not true-to-life, if it is used.	**2** **The writing tends to be choppy or rambling. Awkward construction often forces the reader to slow down and reread. The writing has** • large portions of the text that are difficult to follow or read aloud. • sentence patterns that are monotonous (e.g., subject-verb or subject-verb-object). • a large number of awkward, choppy, or rambling sentence structures.	**1** **The writing is difficult to follow or to read aloud. Sentences tend to be choppy,incomplete, rambling, or just very awkward. The writing has** • sentences that may be hard to read aloud easily. • confusing word order that often makes the meaning hard to follow. • sentence patterns that frequently make meaning unclear. • sentences that are fragmented, confusing, choppy, or rambling on and on.

Scoring Guide for Oregon's Writing Assessment

Word Choice **Selecting functional, precise and descriptive words appropriate for audience and purpose**		
6 **Words communicate the intended message in an exceptionally interesting, accurate and natural way. The writer uses a rich, broad range of words that have been carefully chosen and thoughtfully placed. The writing has** • accurate, powerful and specific words; word choices make the writing interesting and lively. • fresh, original expression; if slang is used, it is for a reason and works very well. • vocabulary that has variety and gets noticed but is also natural and doesn't seem to be trying to impress the reader. • ordinary words used in an unusual way. • words that create strong pictures in the reader's mind; metaphors and similes may be used.	**5** **Words communicate the intended message in an interesting, accurate, and natural way. The writer uses a broad range of words that have been carefully chosen and thoughtfully placed. The writing has** • accurate, specific words; word choices make the writing more interesting and lively. • fresh, clear expression; if slang is used, it is for a reason and works well. • vocabulary that may have variety and get noticed but is also natural and doesn't seem to be trying to impress the reader. • ordinary words used in an unusual way. • words that create clear pictures in the reader's mind; metaphors and similes may be used.	**4** **Words communicate the intended message. The writer uses a variety of words that work and are appropriate for the topic, audience and purpose. The writing has** • words that work but do not necessarily make the writing more interesting and lively. • expression that works; however, slang, if used, does not always seem to match the purpose or seem effective. • some attempts at colorful language; however, they may occasionally seem overdone. • rare experiments with language; however, the writing may have some especially good moments, and it generally avoids clichés.
3 **Language is ordinary. The writer does not use a variety of words, producing a sort of "generic" paper with commonly used words and phrases. Words may be too technical or loaded with jargon. The writing has** • words that work, but that are rarely interesting. • expression that seems ordinary and general; any slang is used for a reason and is effective. • words that are accurate for the most part, although misused words may sometimes appear. • attempts at colorful language that do not fit or seem natural; they seem forced or trying to impress. • too many clichés and overused expressions. • overuse or ineffective use of technical jargon.	**2** **The language is monotonous and/or misused, taking away from the meaning and impact. The writing has** • words that are flat or not specific enough. • words or expressions that are either so common or used so often that they detract from the message. • images that don't work because they are not clear or are absent altogether.	**1** **The writing shows a limited vocabulary, or is so filled with words not used correctly that the meaning is unclear. Only the most general idea comes through because the language is not specific enough. The writing has** • general, vague words that do not make the point. • a small set of words used over and over. • words that simply do not work; they seem too general or just plain wrong.

Scoring Guide for Oregon’s Writing Assessment

Conventions **Demonstrating knowledge of spelling, grammar, punctuation, capitalization, usage, paragraphing**		
6 **The writing demonstrates mastery of a variety of standard conventions, even in complex and less common situations. Errors, i f any,are not obvious or significant. The writing has** • correct use of punctuation, including commas, semicolons, apostrophes and colons, in a variety of situations to add meaning. • correct spelling, even of difficult words. • paragraphing that strengthens the impact and organization. • correct capitalization. • correct grammar and usage that contribute to clarity and style. • skill in using a wide range of conventions in a sufficiently long and complex piece. • little or no need for editing.	**5** **The writing demonstrates strong control of standard conventions which effectively contribute to the message. Errors are so few and so minor that they do not distract the reader. The writing has** • correct grammar and usage. • sound paragraphing. • effective use of punctuation. • correct spelling, even of difficult words. • few capitalization errors. • skill in using a wide range of conventions in a sufficiently long and complex piece. • little need for editing.	**4** **The writing demonstrates competent handling of standard conventions. Minor errors are distracting but not confusing. The writing has** • correct end-of-sentence punctuation; minor and very few or no instances of confusion with commas, semi-colons, apostrophes or colons. • common or key words spelled correctly. • paragraph breaks that are logically placed. • correct capitalization; errors, if any, are in uncommon cases. • occasionally incorrect grammar and usage; problems do not confuse or change the meaning. • a need for some minor editing.
3 **The writing shows a limited control of standard conventions. Errors begin to interfere with readability. The writing has** • errors in grammar, usage, and capitalization that do not block meaning but do distract the reader. • paragraphs that sometimes run together or begin at ineffective points. • end-of-sentence punctuation that is usually correct, but internal punctuation contains frequent errors. • spelling errors that distract the reader; misspelling of common words sometimes occurs. • some control over basic conventions, but the text is too simple or too short to show mastery. • a significant need for editing.	**2** **The writing shows little understanding of standard conventions. Errors often distract and confuse the reader, requiring the reader to reread passages. The writing has** • many places where punctuation is left out or incorrect. • frequent spelling errors, even of common words. • random paragraph indentations or none at all. • many capitalization errors, including sentence beginnings and names. • errors in grammar and usage that confuse the reader or change the meaning or are inappropriate for audience and purpose. • a need for major revisions and corrections.	**1** **Numerous errors in conventions repeatedly distract the reader and make the writing difficult to read. The writing has** • very limited skill in using conventions. • punctuation (including ends of sentences) that tends to be omitted, haphazard, or incorrect. • frequent spelling errors that significantly interfere with readability. • paragraphing that may be irregular or absent. • capitalization that appears to be random. • a need for extensive editing.

Scoring Guide for Oregon’s Writing Assessmentw

Citing Sources Use only on classroom assignments requiring research **Indicating the sources of information presented, including all ideas, statements, quotes and statistics that are taken from sources and that are not common knowledge**		
6 **The writing demonstrates exceptionally strong commitment to the quality and significance of research and the accuracy of the written document. Documentation is used to avoid plagiarism and to enable the reader to judge how believable or important a piece of information is by checking the source. The writer has** • acknowledged borrowed material by introducing the quotation or paraphrase with the name of the authority. • punctuated all quoted materials; errors, if any, are minor. • paraphrased material by rewriting it using writer’s style and language. • provided specific in-text documentation for each borrowed item. • provided a bibliography page listing every source cited in the paper; omitted sources that were consulted but not used.	**5** **The writing demonstrates a strong commitment to the quality and significance of research and the accuracy of the written document. Documentation is used to avoid plagiarism and to enable the reader to judge how believable or important a piece of information is by checking the source. Errors are so few and so minor that the reader can easily skim right over them unless specifically searching for them. The writer has** • acknowledged borrowed material by introducing the quotation or paraphrase with the name of the authority; key phrases are directly quoted so as to give full credit where credit is due. • punctuated all quoted materials; errors are minor. • paraphrased material by rewriting using writer’s style and language. • provided specific in-text documentation for borrowed material. • provided a bibliography page listing every source cited in the paper; omitted sources that were consulted but not used.	**4** **The writing demonstrates a commitment to the quality and significance of re search and the accuracy of the written document. Documentation is used to avoid plagiarism and to enable the reader to judge how believable or important a piece of information is by checking the source. Minor errors, while perhaps noticeable, do not blatantly violate the rules of documentation. The writer has** • acknowledged borrowed material by sometimes introducing the quotation or paraphrase with the name of the authority. • punctuated all quoted materials; errors, while noticeable, do not impede understanding. • paraphrased material by rewriting using writer’s style and language. • provided in-text documentation for most borrowed material. • provided a bibliography page listing every source cited in the paper; included sources that were consulted but not used.
3 **The writing demonstrates a limited commitment to the quality and significance of research and the accuracy of the written document. Documentation is sometimes used to avoid plagiarism and to enable the reader to judge how believable or important a piece of information is by checking the source. Errors begin to violate the rules of documentation. The writer has** • enclosed quoted materials within quotation marks; however, incorrectly used commas, colons, semicolons, question marks or exclamation marks that are part of the quoted material. • included paraphrased material that is not properly documented. • paraphrased material by simply rearranging sentence patterns.	**2** **The writing demonstrates little commitment to the quality and significance of research and the accuracy of the written document. Frequent errors in documentation result in instances of plagiarism and often do not enable the reader to check the source. The writer has** • enclosed quoted materials within quotation marks; however, incorrectly used commas, colons, semicolons, question marks or exclamation marks that are part of the quoted material. • attempted paraphrasing but included words that should be enclosed by quotation marks or rephrased into the writer’s language and style. • altered the essential ideas of the source. • included citations that incorrectly identify reference sources.	**1** **The writing demonstrates disregard for the conventions of research writing. Lack of proper documentation result in plagiarism and do not enable the reader to check the source. The writer has** • borrowed abundantly from an original source, even to the point of retaining the essential wording. • no citations that credit source material. • included words or ideas from a source without providing quotation marks. • included no bibliography page listing sources that were used.

Source: From *Writing Scoring Guide: Middle School Student Version* (pp. 1–7), by Oregon Department of Education, 1996. Salem: Office of Assessment and Evaluation, author. Reprinted by permission.

Basic Statistical Concepts

It is necessary for you to have an understanding of a few basic statistical concepts to better understand the results of your classroom assessments, to better summarize assessment results when you grade students, to interpret your students' norm-referenced test scores, to understand the basic data in published test manuals, and to understand assessment summary reports provided by your school district or state. This appendix focuses on concepts rather than on computations. However, the computations of certain statistical indices are illustrated so you will understand the origin of their numerical values.

Although you may believe that mathematics or computations are your weak suit, you should not shy away from learning the few techniques shown in this appendix. With the availability of inexpensive calculators, computations become simple and accurate with only a little practice. Some computerized gradebook programs will make a few calculations. You should also buy an inexpensive scientific calculator that has a "correlation" or "*r*" function. Such a calculator will allow you to enter the scores from your assessments and painlessly carry out calculations for all of the statistical indices in this appendix.

Statistical methods are techniques to summarize scores so that you may better understand how a group of students has performed and how well an individual student has performed relative to others in the group. A **statistical index** (or **statistic**) is a summary number that concisely captures a specific feature of a group of scores. For example, measures of central tendency focus on an average or typical score for a group. Measures of variability focus on quantifying the extent to which students' scores differ from one another. This appendix presents four categories of statistical methods that you will find most useful in understanding test scores and other assessment results: (1) distribution of scores, (2) typical or average score, (3) variability of scores, and (4) degree to which two sets of scores are correlated.

DESCRIBING DISTRIBUTIONS OF TEST SCORES

Suppose the scores shown in Figure I.1 are scores of our students on two tests you gave. The arrangement of the scores in the table is similar to how they might be arranged in your gradebook: Students' names are arranged alphabetically with their mark next to their names. This arrangement does not make it easy to answer such questions as:

- How many students in the class have similar scores?
- What scores do most students obtain?
- Are the scores widely scattered along the score scale, or do they bunch together?
- Does the pattern of scores in the class appear unusual in some way? Or are they as expected?
- Does any student score unusually higher or lower than his or her classmates?

Ranking Scores

One simple way to begin answering questions such as these is to rank the scores. Most people know how to do this already. To rank the scores, *order them from largest to smallest*. The largest score is assigned a rank of 1; the next largest, a rank of 2; and so on, down to the smallest score. In this way all of the raw scores (marks) are transformed into ranks.

Figure I.2 demonstrates the procedure for scores from Test 1. Notice what is done when students have the same score. In this case they are tied for the ranks. The tie is resolved by awarding each of the persons whose scores are tied the average of the ranks for which they are tied. For example, four students have a score

From Appendix I of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FIGURE I.1 List of students in a class and their scores on two tests.

Name	Test 1	Test 2
1. Anthony	89	94
2. Ashley	75	68
3. Blake	74	72
4. Chad	84	77
5. Donald	56	66
6. Edward	80	68
7. Festina	66	68
8. George	86	73
9. Harriet	68	73
10. Irene	98	86
11. Jesse	65	78
12. Katherine	44	60
13. Lorraine	45	53
14. Marya	61	75
15. Nancy	75	76
16. Oprah	68	54
17. Peter	55	53
18. Quincy	70	68
19. Robert	69	65
20. Sally	60	47
21. Tina	73	74
22. Ula	75	88
23. Veronica	71	73
24. Wallace	43	61
25. William	83	87
26. Xavier	95	83
27. Yvonne	96	85
28. Zena	75	70

FIGURE I.2 Rank order of students from Figure I.1 according to their scores on Test 1.

Name	Test 1	Rank	
Irene	98	1	
Yvonne	96	2	
Xavier	95	3	
Anthony	89	4	
George	86	5	
Chad	84	6	
William	83	7	
Edward	80	8	
Zena	75	10.5	Four scores tied for ranks 9, 10, 11, and 12
Ula	75	10.5	
Nancy	75	10.5	
Ashley	75	10.5	
Blake	74	13	
Tina	73	14	
Veronica	71	15	
Quincy	70	16	
Robert	69	17	
Oprah	68	18.5	Two scores tied for ranks 18 and 19
Harriet	68	18.5	
Festina	66	20	
Jesse	65	21	
Marya	61	22	
Sally	60	23	
Donald	56	24	
Peter	55	25	
Lorraine	45	26	
Katherine	44	27	
Wallace	43	28	

of 75 and thus are tied for ranks 9, 10, 11, and 12. Rather than arbitrarily awarding one person a rank of 9, and another a rank of 10, and so on, each person is awarded the average of the tied ranks, that is:

$$\frac{9 + 10 + 11 + 12}{4} = 10.5$$

A simple ranked list of scores helps you answer some basic questions about how well your class performed on a test. The list shows quickly the highest and lowest scores. It shows how the scores are spread out and which scores occur most often. This ranked list may be all you need to understand how your students performed on a test. However, ranked lists are not easily understood if the number of students is very large: for example, the score of all fourth graders in the school district or in the state. A better way to organize the scores in such cases is discussed later.

Interpretations of simple ranks of this sort depend on the number of students in the group. For example, suppose I told you that of all the classes in testing and measurement I have taught, your class ranked second. You might be proud as a group until I also told you that I have taught only one other class. Adding another 13 classes might result in your class's rank dropping, say, from second to 15th: Although the class's rank has changed from second to 15th, its relative position—dead last—has not changed. The point is that a student's rank cannot be fully interpreted without knowing the number of other students being ranked. This problem is largely overcome by using *percentile ranks* (see Chapter 16). We show you how to calculate percentile ranks in Figure I.8.

Stem-and-Leaf Displays

A simple way to organize a large group of scores is to prepare a stem-and-leaf display. Figure I.3 illustrates the procedure for the scores of the 28 students in Figure I.1 for each of the two tests. The "stem" is the tens' digit and the "leaves" are the ones' digits of the score. For example, consider the scores 80, 83, 84, 86, and 89 from Test 1. The tens' digit is 8 and is written in the stem column. The ones' digits 0, 3, 4, 6, and 9 are the leaves and are written in the row to the right of the 8.

FIGURE I.3 **Stem-and-leaf display of the distribution of the students' scores from Figure I.1.**

Test 1			Test 2		
Stem	**Leaves**	**Frequency**	**Stem**	**Leaves**	**Frequency**
0			0		
1			1		
2			2		
3			3		
4	3 4 5	3	4	7	1
5	5 6	2	5	3 3 4	3
6	0 1 5 6 8 8 9	7	6	0 1 5 6 8 8 8 8	8
7	0 1 3 4 5 5 5 5	8	7	0 2 3 3 3 4 5 5 6 7 8	10
8	0 3 4 6 9	5	8	3 5 6 7 8	5
9	5 6 8	3	9	4	1
		$N = 28$			$N = 28$

The stem-and-leaf display has the advantage of showing how the entire group of scores is distributed along the score scale when they are grouped together by intervals of 10. That is, it organizes the scores into the groupings of 40s, 50s, 60s, 70s, 80s, and 90s. With the ones' digits displayed, you can easily "reconstitute" individual values of the scores. This is useful if you need to make future calculations. In the Frequency column, the number of scores is written in each row.

Notice that tens' digits in the stem column (0, 1, 2, etc.) are ordered from lowest to highest. When you turn the page on its side, the display is a type of graph: The length of the "leaves" row is proportional to the frequency of the scores.

The scores in Figure I.3 are grouped into interval widths of 10. You could also group the scores into narrower intervals, say five digits wide, as shown in Figure I.4. The stem 4 represents the scores 40, 41, 42, 43, and 44; the stem 4* represents the scores 45, 46, 47, 48, and 49. Figures I.3 and I.4 contain the same information, but are organized slightly differently. Notice, too, that you can easily construct a ranked list from a stem-and-leaf display, because the individual score values are easily recovered from the display.

Frequency Distributions

When the number of scores to be organized is large, ranked lists and stem-and-leaf displays are cumbersome.

FIGURE I.4 **Stem-and-leaf display of the scores from Figure I.1 when the internal width equals 5.**

Test 1			Test 2		
Stem	**Leaves**	**Frequency**	**Stem**	**Leaves**	**Frequency**
0			0		
0*			0*		
1			1		
1*			1*		
2			2		
2*			2*		
3			3		
3*			3*		
4	3 4	2	4		
4*	5	1	4*	7	1
5		0	5	3 3 4	3
5*	5 6	2	5*		0
6	0 1	2	6	0 1	2
6*	5 6 8 8 9	5	6*	5 6 8 8 8 8	6
7	0 1 3 4	4	7	0 2 3 3 3 4	6
7*	5 5 5 5	4	7*	5 6 7 8	4
8	0 3 4	3	8	3	1
8*	6 9	2	8*	5 6 7 8	4
9		0	9	4	1
9*	5 6 8	3	9*		0
		$N = 28$			$N = 28$

In such cases, the collection of scores is organized into a figure called a **frequency distribution.** This figure shows the number of persons obtaining various scores. Figure I.5 shows frequency distributions for the two tests.

Notice that the figure shows scores grouped into intervals of 5 points on the score scale: 95–99, 90–94, 85–89, and so on. Grouping scores into intervals is a common practice when the students' scores span a wide range of values. The advantage is that the table shows the distribution of scores in a more compact space. The number of intervals is set at some convenient value, say 10 or 12. A common practice is to make the width of the interval an odd number, because then the midpoint of the interval is a whole number. Whole-number midpoints are desirable when the information in the table is to be used to construct a graph or for later calculations. The midpoints of each interval are shown in Figure I.5. Often the midpoints are not presented when the table can be interpreted without that information. Similarly, the Tally column is seldom shown in a finished table. Its only purpose is to make it easier and more accurate to count the scores in each interval. At the bottom of the frequency column you should record the sum of the frequencies. This is *N*, the total number of scores in the collection.

You can calculate the width of the interval to use as follows. Subtract the lowest score in the group from the highest score. Divide this difference by 12. The interval width to use is the nearest odd number to this quotient. For example, for Test 1, the highest score is 98 and the lowest is 43. Thus, (98 − 43) ÷ 12 is 4.56, and the nearest odd number is 5. For Test 2, this calculation is 3.91 and the nearest odd number is 3. However, if you want to compare the distributions of Tests 1 and 2, it is best to use the same interval width. Thus, we have used a width of 5 for each test distribution. This illustrates that there are no hard and fast rules for fixing interval widths.

To make the table, it is best to make the lower limit of the interval a multiple of the interval width. This makes it easier to construct the table. Thus the lower limit of the highest interval in Figure I.5 is 95, the next is 90, next is 85, and so on. Be sure that the highest interval contains the highest score. You need not continue the intervals below the interval containing the lowest score. Thus the lowest interval in Figure I.5 is 40–44.

The **grouped frequency distribution,** as Figure I.5 is called, provides the same convenient summary of the distribution of scores as the stem-and-leaf display. However, unlike the stem-and-leaf display, information about the specific numerical values of the scores in each interval is lost: Only the frequency of the scores falling into the interval is recorded. Unlike the stem-and-leaf display, however, the frequency distribution table can summarize large collections of scores in a compact, easy-to-interpret format.

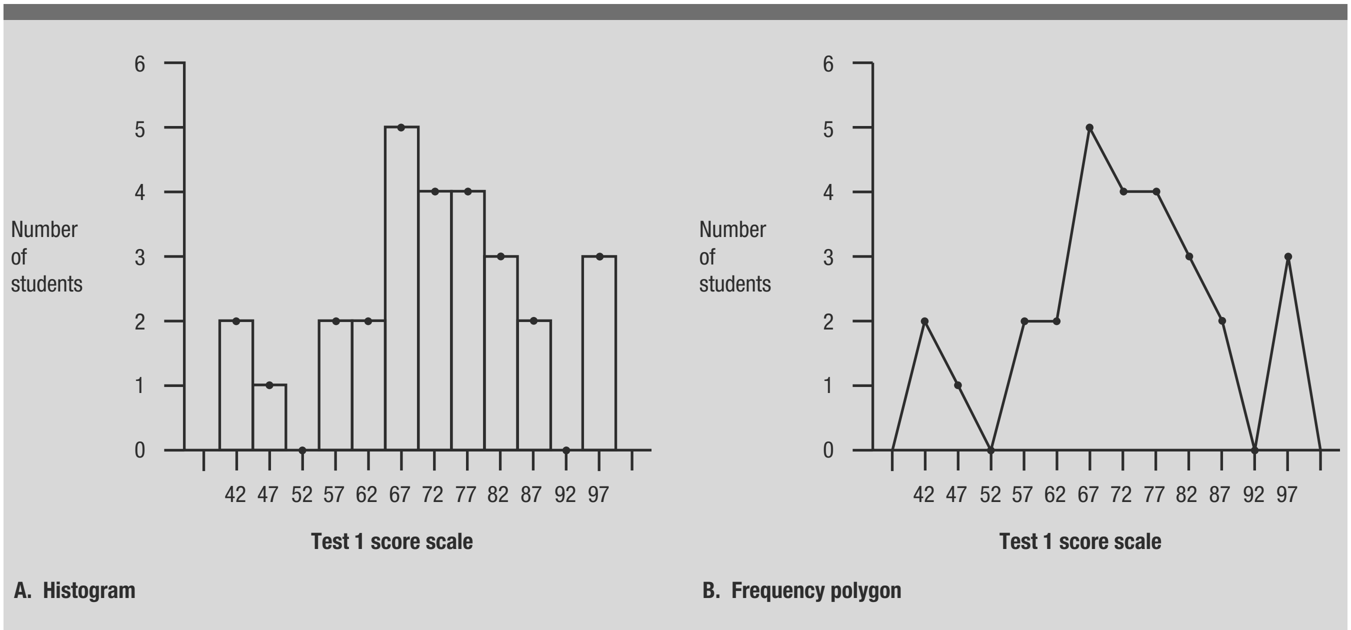
Frequency Polygons and Histograms

Frequency distributions are often graphed because graphs permit an increased understanding of the distribution of scores. Two common types of graphs of frequency distribution are the histogram and frequency polygon. For both, a scale of score values is marked off on a horizontal axis. The **histogram** (sometimes called a *bar graph*) represents the frequency of each score by a rectangle. The height of each rectangle is made equal to (or proportional to) the frequency of the corresponding score. Figure I.6A shows a histogram for the Test 1 scores of Figure I.5. A **frequency polygon** for these same scores is shown in Figure I.6B. A dot is made directly above the score-value to indicate the frequency. (If no one has obtained a particular score-value, the dot is

FIGURE I.5 Frequency distributions of the scores in Figure I.1. (Interval widths equal 5.)

Test 1				Test 2			
Interval	Tally	Midpoint	Frequency	Interval	Tally	Midpoint	Frequency
95–99	|||	97	3	95–99		97	0
90–94		92	0	90–94			92	1
85–89	||	87	2	85–89	|||	87	4
80–84	|||	82	3	80–84			82	1
75–79	||||	77	4	75–79	||||	77	4
70–74	||||	72	4	70–74	𝍸		72	6
65–69	𝍸	67	5	65–69	𝍸		67	6
60–64	||	62	2	60–64	||	62	2
55–59	||	57	2	55–59		57	0
50–54		52	0	50–54	|||	52	3
45–49			47	1	45–49			47	1
40–44	||	42	2	40–44		42	0
			28				28

FIGURE I.6 Histogram and frequency polygon for the scores of Test 1 from Figure I.5.



A. Histogram

B. Frequency polygon

made at 0.) The dots are then connected with straight lines to make the polygon.

A graph communicates in an easy manner the shape or form of a frequency distribution. Using the names of these shapes is a compact way of describing how the scores are distributed.

Figure I.7 shows a variety of distributional forms, their corresponding names, and examples of measurement situations that might give rise to them.

Frequency Distributions and Graphs

The illustrations of Figure I.7 are idealized and do not represent actual distributions. Nevertheless, it is helpful to have a mental picture of these distributional forms because, in practice, actual test score distributions resemble the ideal forms at least roughly. Test manuals and school and state reports often describe score distributions using the terms.

Score Distribution Shape Depends on Both the Test Taker and the Test

The shape of a score distribution reflects the characteristics of the test as well as the ability of the group being tested. There is no single "natural" or "normal" shape toward which the test scores of a given group of students tends. A test composed of items that are not too difficult and not too easy for a particular group is likely to result in distributions of scores similar to those illustrated by A, B, or C of Figure I.7. This same group, with the same ability, could take a test in the same subject made up of items that few persons could answer correctly or a test made up of "easy" items. In these latter cases, skewed distributions (F or G) might result. It is not accurate, therefore, to come to a conclusion about the *underlying ability* distribution of a group of students by examining only the shape of the distribution of observed test scores. The characteristics of the test the group took also need to be made a part of the decision.

Choice Between Histogram or Polygon

For many classroom purposes, you could use either a polygon or a histogram; the choice between them is rather arbitrary. The polygon emphasizes the continuous nature of the attribute that underlies the scores you see on the test; the histogram emphasizes the discrete nature. Although observed test *scores* usually are discrete whole numbers (0, 1, 2, . . .), the underlying characteristic a test is designed to measure is often thought of as continuous rather than discrete.

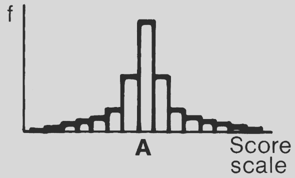
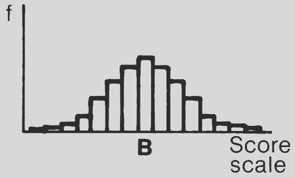
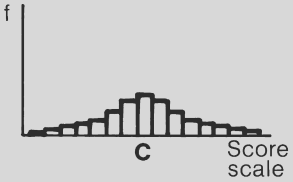
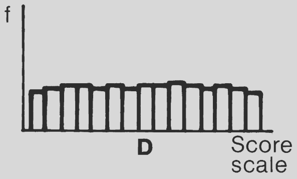
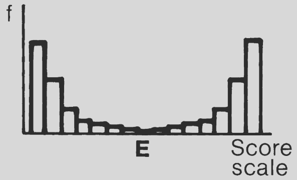
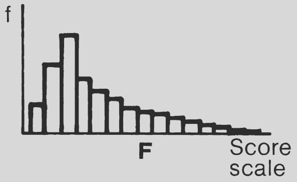
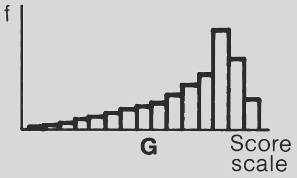
Comparing Two Distributions

It is sometimes useful to compare two or more frequency distributions by graphing on the same axes using polygons, rather than histograms. A graph could compare, for example, a class of students before and after instruction, or it could compare two different classes of students. Such a graph would display the forms of the distributions, the variability or disbursement of scores, and the place(s) along the score scale where the scores tend to cluster.

Calculating Percentile Ranks

Percentile ranks tell the percentage of the scores in a distribution that are below a particular point on the score scale. We explained the concept of percentile ranks in Chapter 16. We show how to calculate percentile ranks in Figure I.8.

FIGURE I.7 Histograms showing various forms of frequency distributions.

	Histogram	Description of distribution form	Examples of when such shapes might occur
Symmetrical		**A.** Unimodal, symmetrical but relatively peaked.	A distribution of scores on an arithmetic test of medium difficulty.
		**B.** Unimodal, symmetrical with moderate degree of peakedness.	A distribution of scores on an arithmetic test of medium difficulty.
		**C.** Unimodal, symmetrical but relatively flat.	A distribution of scores on an arithmetic test of medium difficulty.
		**D.** Rectangular or uniform.	A distribution of monthly incidence of infant mortality in a large modern hospital.
		**E.** Bimodal and symmetrical or U-shaped.	A distribution of ages at time of death of pedestrians killed by automobiles.
Skewed		**F.** Positively skewed or skewed to right.	A distribution of scores on a "hard" arithmetic test.
		**G.** Negatively skewed or skewed to left.	A distribution of scores on an "easy" arithmetic test.

Source: Adapted from *Elementary Statistical Methods in Psychology and Education* (2nd ed., pp. 30–31), by P. J. Blommers and R. A. Forsyth, 1977, Boston: Houghton-Mifflin. Adapted by permission of the author.

FIGURE I.8 Example of how to calculate percentile ranks for a class of 25 students.

Raw	Tally	Frequency	Cumulative frequency	*Percentile rank* $PR = \frac{\frac{1}{2}\left[\text{number of persons having the score}\right] + \left[\text{number of persons below the score}\right]}{\text{total number of persons}} \times 100$
36	/	1	25	$98 = \frac{.5 + 24}{25} \times 100$
35		0	24	96
34		0	24	96
33		0	24	$96 = \frac{0 + 24}{25} \times 100$
32	/	1	24	$94 = \frac{.5 + 23}{25} \times 100$
31	/	1	23	$90 = \frac{.5 + 22}{25} \times 100$
30		0	22	$88 = \frac{0 + 22}{25} \times 100$
29	//	2	22	$84 = \frac{1 + 20}{25} \times 100$
28	////	4	20	$72 = \frac{2 + 16}{25} \times 100$
27	𝍸	5	16	$54 = \frac{2.5 + 11}{25} \times 100$
26	𝍸 /	6	11	$32 = \frac{3 + 5}{25} \times 100$
25	//	2	5	$16 = \frac{1 + 3}{25} \times 100$
24	/	1	3	$10 = \frac{.5 + 2}{25} \times 100$
23		0	2	8
22		0	2	$8 = \frac{0 + 2}{25} \times 100$
21	/	1	2	$6 = \frac{.5 + 1}{25} \times 100$
20		0	1	4
19		0	1	4
18		0	1	4
17		0	1	4
16		0	1	4
15		0	1	$4 = \frac{0 + 1}{25} \times 100$
14	/	1 N = 25	1	$2 = \frac{.5 + 0}{25} \times 100$

Step-By-Step

1. List the possible scores in descending order (Column 1). (You may group the scores into intervals if you wish.)
2. Ta lly the number of students attaining each score (Column 2).
3. Sum the number of students attaining each score (Column 3).
4. Add the frequencies consecutively, starting at the bottom of the column with the lowest score. Place each consecutive sum in the cumulative frequency column (Column 4). E.g., 0 + 1 = 1, . . . , 2 + 1 = 3, 3 + 2 = 5, etc.
5. Calculate the percentile rank of each score (Column 5). Below is an example for the score 27.
   - (a) Calculate one-half of the frequency of the score (1/2 × 5 = 2.5).
   - (b) Add the result in (a) to the cumulative frequency just below the score (2.5 + 11 = 13.5).
   - (c) Divide the result in (b) by the total number of scores (13.5 ÷ 25 = .54).
   - (d) Multiply the result in (c) by 100 (.54 × 100 = 54).

MEASURES OF CENTRAL TENDENCY

It is quite common when interpreting assessment results to speak of the "average score," as we speak of the "average student," "being above average in spelling," or "of average intelligence."

There are many ways to define averages, but we describe only three: mode, mean, and median. The **mode** is the score that occurs most frequently in relation to other scores in the collection. Thus, the modal score is average in the sense of being most popular or most probable in the group. The **mean,** or more precisely the *arithmetic mean,* is found by summing the scores and dividing by their number. Thus, the mean is the average that takes into account all of the scores and is the "center of gravity" of the collection. The **median** is the score point that divides the score scale so 50% of the scores in the collection are above it and 50% are below it. This makes the median a typical score in the sense of coming nearest in the aggregate to all the scores.

Mode

You find the mode by listing the scores and identifying the most frequently occurring. In Figure I.9, the mode of the Test 1 distribution is 75, and the mode of the Test 2 distribution is 68. You could identify the mode from either a stem-and-leaf display or a frequency distribution.

FIGURE I.9 Scores on Test 1 and Test 2 ranked separately and showing measures of central tendency.

Test 1		Test 2	
98		94	
96		88	
95		87	
89		86	
86		85	
84		83	
83		78	
80		77	
75	Mode = 75	76	
75		75	
75		74	
75		73	
74		73	
73	← Median = 72	73	←Median = 72.5
71		72	
70		70	
69		68	Mode = 68
68		68	
68		68	
66		68	
65		66	
61		65	
60		61	
56		60	
55		54	
45		53	
44		53	
43		47	
$\Sigma X = 1{,}999$		$\Sigma X = 1{,}995$	
$M = (1{,}999) \div 28 = 71.4$		$M = (1{,}995) \div 28 = 71.3$	

If in one distribution two scores occur with approximately equal frequency, there are two modes. Such a distribution is said to be **bimodal**. A distribution with one mode is **unimodal.**

The mode is the point on the score scale where a large number of scores in a distribution are located. If there is more than one mode, there are concentrations of scores at more than one score level. You should note that a distribution may not have a mode. For example, the uniform distribution in Figure I.7D does not have a mode.

Mean

To calculate the mean, add the scores and divide by their number. The formula is

$$M = \frac{\textit{Sum of all the scores}}{\textit{Total number of scores}}$$
$$= \frac{\Sigma X}{N}$$

where M represents the mean, N represents the total number of scores involved, and Σ represents "sum of." The means of the Test 1 and Test 2 scores in Figure I.9 are 71.4 and 71.3, respectively.

An important property of the mean is that its value is affected by every one of the scores in the collection, because the sum on which it is based includes every score. When you want an average that focuses on the total rather than the typical or most frequent, choose the mean. The mean reflects the highest and lowest scores, whereas the mode reflects only the most frequent. This influence of extremely high or low scores may be undesirable because such scores are not typical scores for the distribution. The median is preferred when you want an average to focus on typical performance and to be uninfluenced by extremely high or extremely low scores.

Median

A simple way to calculate the median is to arrange the scores by rank, and then count to the point on the score scale that has the same number of scores above it as below it. If there is an even number of scores in the collection, the median is halfway between the two middle scores. If there is an odd number of scores, the median is the middle score.

In Figure I.9, the median for Test 1 is 72; the median for Test 2 is 72.5. Notice that the median does not have to be a score that any person has attained. This is so because the median is a point on the score scale that divides the distribution into halves. The mean also need not be a score anyone attained; however, the mode must be a score that many persons attained.

Because the median separates the distribution into two halves, it is also the *50th percentile*. Further, it does not sum up all of the scores. As a result, its value is not affected by extremely high or low scores (as the mean is). The median is the average to use when you do not want an average that is sensitive to such extreme scores.

MEASURES OF VARIABILITY

Although averages summarize the central tendency of a group of scores, they do not summarize how the scores spread out over the score scale. For example, the mean reading test scores of two seventh-grade classes may be 75. However, in one class the scores may range widely from 55 to 95, while in the other the scores may range only from 70 to 80. Obviously, the students in the latter class are more nearly alike in their reading achievement than the students in the former class. You will need to cater to more widely different reading levels when teaching the former class than when teaching the latter.

This section describes three measures of the spread or variability of a set of scores: the range, the interquartile range, and the standard deviation.

Range (R)

The **range** is a simple index of spread. It is the difference between the highest and lowest scores in the set. For the two tests in Figure I.9, the range is 55 for Test 1 (98 − 43 = 55) and 47 for Test 2 (94 − 47 = 47). Although for either test the range is relatively large, it is smaller for Test 2, showing the scores are spread over a smaller part of the score scale. The procedure may be summarized as follows:

$$R = \textit{highest score} - \textit{lowest score}$$

A weakness of the range as an index of variability is that it is based on only two scores. It ignores the scores between the highest and lowest scores. Another problem with the range is that a change in either the highest or lowest score in the set can radically alter its value.

Interquartile Range (IR)

The **interquartile range** describes the spread of the middle 50% of the scores. It is the difference between the third and the first quartiles. **Quartiles** are points that divide the group of scores into quarters. The first quartile(Q_1) is the point *below which* the lowest 25% of the students score. The third quartile(Q_3) is the point *above which* the highest 25% of the students score. The second quartile(Q_2) is the median.

To obtain the interquartile range, you first order the scores and proceed similarly to calculating the median. That is, count down from the highest score 25% of the scores to locateQ_3 and up from the lowest score 25% to calculate Q_1. The interquartile range is the difference between these two values:[1]

$$IR = Q_3 - Q_1$$

In Figure I.9, Q_1 = 83 and Q_1 = 61 for Test 1. That is, for Test 1 the 75th percentile is 83 and the 25th percentile is 61. The IR = [83 – 61] = 22 for this test. Thus, the middle 50% of the scores on Test 1 have a 22-point spread. For Test 2, Q_3 = 73, Q_1 = 65, and IR = [73 – 65] = 8. The middle 50% of the students have only an 8-point spread on Test 2.

Standard Deviation (SD)

The most frequently used index of variability is the **standard deviation.** Large numerical values of this index indicate that the scores are spread out away from the mean. Small values indicate that the scores tend to cluster near the mean. The standard deviation is the average amount by which the scores differ from the mean score.[2] In some test reports the squared standard deviation (SD^2), or *variance*, is used.

The definitional formula for the standard deviation is:

$$SD = \sqrt{\frac{\sum (X - M)^2}{N}}$$

$$= \sqrt{\frac{\textit{Sum of the Squared deviations from the mean}}{\textit{total number of scores}}}$$

Many inexpensive scientific calculators and microcomputer programs have procedures for calculating the standard deviation. You should use one of these to calculate SD. If you want to calculate the standard deviation using a calculator that does not have this procedure built in, follow these steps:

1. First arrange the scores into a frequency distribution, as in Figure I.5 and reproduced in Figure I.10.
2. Apply a computational formula such as the one shown here. (You can find other computational formulas in an applied statistics text.)

$$SD = \sqrt{\frac{\Sigma f(X^2)}{N} - M^2}$$

$$= \sqrt{\frac{\begin{array}{c}\textit{sum of the product of the square}\\ \textit{of each score and its frequency}\end{array}}{\textit{total number}} - \left[\begin{array}{c}\textit{square of}\\ \textit{the mean}\end{array}\right]}$$

[1] Some books divide the interquartile range by 2 to obtain the *semi- interquartile range (SIR).* This value indicates the approximate distance you would need to move on the score scale above and below the median to encompass the middle 50% of the scores.

[2] This is not strictly correct, but as a practical matter little interpretive harm regarding assessment results is done by thinking of the standard deviation in this way.

FIGURE I.10 Computing the standard deviation of Test 1 scores after they are organized into a frequency distribution.

Score Interval	Midpoint	Frequency (f)	Step 1 (X^2)	Step 2 f(X^2)
95–99	97	3	9,409	28,227
90–94	92	0	8,464	0
85–89	87	2	7,569	15,138
80–84	82	3	6,724	20,172
75–79	77	4	5,929	23,716
70–74	72	4	5,184	20,736
65–69	67	5	4,489	22,445
60–64	62	2	3,844	7,688
55–59	57	2	3,249	6,498
50–54	52	0	2,704	0
45–49	47	1	2,209	2,209
40–44	42	2	1,764	3,528
		$n = 28$		150,357

The formula is:

$$SD = \sqrt{\frac{\sum f(X^2) - M^2}{N}} \; (\textit{Note that using the grouped data above, } M = 71.8)$$

Putting the numbers into the formula:

$$SD = \sqrt{\frac{150{,}357}{28} - (71.8)^2}$$

After Steps 4 and 5:

$$SD = \sqrt{5{,}369{,}89 - 5{,}155{,}24}$$

Then Step 6:

$$SD = \sqrt{214.65}$$

Step 7 gives the final result: $SD = 14.65$

This is not as hard to compute as it looks:

Steps	*Symbols*
1. Square each interval midpoint.	1. X^2
2. Multiply each square by its frequency.	2. $f(X^2)$
3. Add together all of these products.	3. $\Sigma f(X^2)$
4. Divide by the total number.	4. $\frac{\Sigma f(X^2)}{N}$
5. Square the mean. (If the mean has not been computed already, you need to compute it.)	5. M^2
6. Subtract the square of the mean from the result found in Step 4. (Stop here if you want only the variance.)	6. $\frac{\Sigma f(X^2)}{N} - M^2$
7. Take the square root of the difference. This is the standard deviation.	7. $\sqrt{\frac{\Sigma f(X^2)}{N} - M^2}$

Figure I.10 illustrates these calculations for Test 1. (Note that in this example, the result obtained from Figure I.10 is not the same result you would obtain if you did not group the scores into intervals. Grouping scores results in some error. However, the result is still useful.)

Calculating Stanines

Stanines are normalized standard scores that tell the location of a raw score in one of nine specific segments of a normal distribution. Stanines were explained in Chapter 16. We show how to calculate stanines in Figure I.11.

THE CORRELATION COEFFICIENT

Calculating the correlation coefficient requires using a calculator or a computer. Some scientific calculators have this capability already built in as a statistical function, so all you need to do is enter the paired scores of students. However, this section illustrates the calculation for those with calculators that do not have the correlation coefficient function built in.

FIGURE I.11 **How to transform raw scores into stanines.**

You may transform any set of scores to stanines by applying the normal curve percentage relationship implied by Figure 17.7. These theoretical percentages are:

Stanine	Percent of scores	Stanine	Percent of scores
9	top 4%	4	next 17%
8	next 7%	3	next 12%
7	next 12%	2	next 7%
6	next 17%	1	bottom 4%
5	middle 20%		

The preferred procedure is to begin assigning stanines at the middle of the score distribution (i.e., assigning stanine = 5 first) and then work toward each end. This procedure helps to make the resulting distribution of stanines more symmetric than if you started at the top or bottom. I illustrate the procedure below using the 25 students' scores shown in percentile rank calibration example given earlier (Figure I.8).

Step-by-step	Results from percentile ranks example	Comments
1. Make a frequency distribution or list the scores in order from high to low.	See PR example, first and third columns.	
2. Locate the median or middle score.	25 scores × $^1/_2$ = 12.5 scores. Therefore, the middle score is 27.	Round the median to a whole number.
3. Use the theoretical percentages to determine how many scores should be assigned a stanine of 5.	20% of 25 = 5 scores.	
4. Assign stanines of 5 to the number of scores calculated in Step 3. (You should include scores just above and below the median if necessary to come as close as possible to the desired number.)	It so happens that in the PR example exactly 5 persons had a score of 27, so that we do not need to look to adjacent values. (See below.)	Remember that *all* equal scores must have the same stanine assigned to them.
5. Working up from the scores assigned stanine 5, use the theoretical percentages to assign scores to stanine categories of 6, 7, 8, and 9. Come as near to the theoretical percentages as possible.		
6. Repeat the procedure for the scores that are below those assigned stanine 5.		
7. It is important that you assign all equal scores the same stanine.		

Scores	Stanines	Actual number	Theoretical number
36	9	1	1
32–35	8	1	2
29–31	7	3	3
28	6	4	4
27	5	5	5
26	4	6	4
24–25	3	3	3
15–23	2	1	2
14	1	1	1

For practical work, you can use the following computational formula:

$$r = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y^2)]}}$$

This formula is illustrated in Figure I.12 with the scores from the two tests in Figure I.1. In this example, Test 1 is symbolized X and Test 2 is symbolized Y. Figure I.13 shows the calculation.

If you have already computed the standard deviations and means of each variable, then the following formula (which is equivalent to the previous equation) will save you some computational labor.

$$r_{xy} = \frac{\dfrac{\Sigma XY}{N} - M_x M_y}{(SD_x)(SD_y)}$$

To illustrate with the data in Figure I.13, for which

$$M_x = 71.39, M_y = 71.25$$
$$SD_x = 14.81, SD_y = 11.57$$

we have:

$$r = \frac{\dfrac{145{,}902}{28} - (71.39)(71.25)}{(14.81)(11.57)}$$

(The slight difference you obtain from these two equations is due to rounding error.)

CALCULATING BASIC STATISTICS WITH THE EXCEL SPREAD SHEET

Many of the statistics in this appendix can be calculated very easily using the Microsoft Excel spreadsheet program. The first figure in this appendix (Figure I.1) shows the scores on two tests for a class of 28 students. We

FIGURE I.12 **Example calculating a correlation coefficient.**

Steps	Symbols	Examples
1. List everyone's pair of scores.	1. X, Y	1. *Blake:* $X = 74$, $Y = 72$
2. Square each score.	2. X^2, Y^2	2. *Blake:* $X^2 = (74)^2 = 5{,}476$ $Y^2 = (72)^2 = 5{,}184$
3. Multiply the scores in each pair.	3. XY	3. *Blake:* $XY = 74 \times 72$ $= 5{,}328$
4. Sum the X, Y, X^2, Y^2, and XY columns.	4. $\Sigma X, \Sigma Y$ $\Sigma X^2\ \Sigma Y^2$ ΣXY	4. $\Sigma X = 1{,}999$, $\Sigma X^2 = 148{,}635$ $\Sigma Y = 1{,}995$, $\Sigma Y^2 = 145{,}761$ $\Sigma XY = 145{,}902$

5. Put the sums into equation.

$$r = \frac{28\ (145{,}902) - (1{,}999)\ (1{,}995)}{\sqrt{[28(148{,}635) - (1{,}999)^2]\,[(145{,}761) - (1995)^2]}}$$

$$= \frac{4{,}085{,}256 - 3{,}988{,}005}{\sqrt{(4{,}161{,}780 - 3{,}996{,}001)\ (4{,}081{,}308 - 3{,}980{,}025)}}$$

$$= \frac{97{,}251}{\sqrt{(165{,}779)\ (101{,}283)}} = \frac{97{,}251}{129{,}578{,}53} = .75$$

FIGURE I.13 **Computing a correlation coefficient between the scores in Figure I.1.**

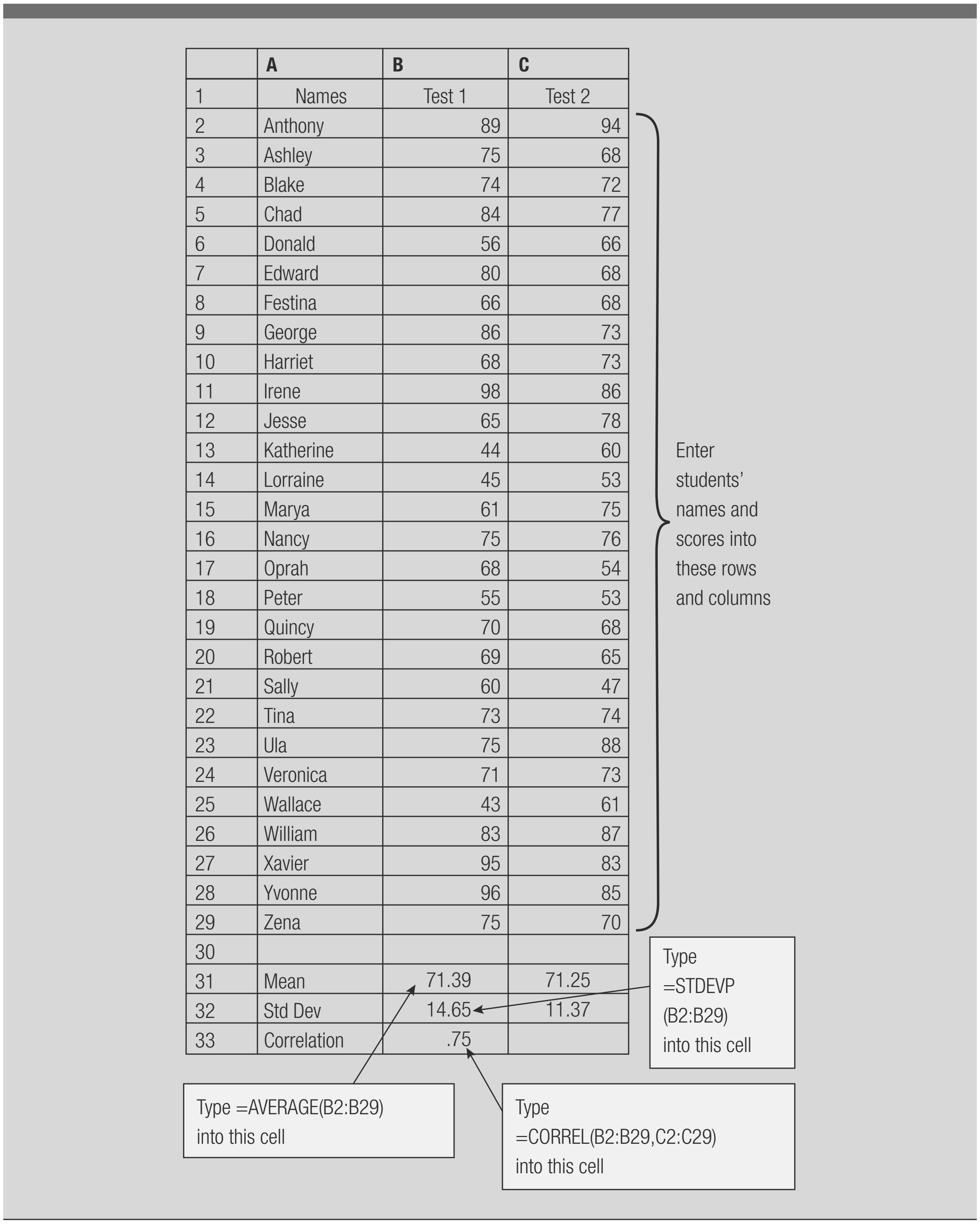
	Test 1		Test 2		Cross Products
Names	***X***	***X***^{2}	***Y***	***Y***^{2}	***XY***
Anthony	89	7,921	94	8,836	8,366
Ashley	75	7,625	68	4,624	5,100
Blake	74	5,476	72	5,184	5,328
Chad	84	7,056	77	5,929	6,468
Donald	56	3,136	66	4,356	3,696
Edward	80	6,400	68	4,624	5,440
Festina	66	4,356	68	4,624	4,488
George	86	7,396	73	5,329	6,278
Harriet	68	4,624	73	5,329	4,964
Irene	98	9,604	86	7,396	8,428
Jesse	65	4,225	78	6,084	5,070
Katherine	44	1,936	60	3,600	2,640
Lorraine	45	2,025	53	2,809	2,385
Marya	61	3,721	75	5,625	4,575
Nancy	75	5,625	76	5,776	5,700
Oprah	68	4,624	54	2,916	3,672
Peter	55	3,025	53	2,809	2,915
Quincy	70	4,900	68	4,624	4,760
Robert	69	4,761	65	4,225	4,485
Sally	60	3,600	47	2,209	2,820
Tina	73	5,329	74	5,476	5,402
Ula	75	5,625	88	7,744	6,600
Veronica	71	5,041	73	5,329	5,183
Wallace	43	1,849	61	3,721	2,623
William	83	6,889	87	7,569	7,221
Xavier	95	9,025	83	6,889	7,885
Yvonne	96	9,216	85	7,225	8,160
Zena	75	7,625	70	4,900	5,250
	1,999	148,635	1,995	145,761	145,902

use the scores in this table to illustrate how to use the Excel program. You need to type students' names and scores into the spreadsheet as shown in the example in Figure I.14. Notice that the rows are labeled with numbers and the columns are labeled with letters. These column letters and row numbers appear automatically and cannot be changed.

To calculate the mean (*M*) of the scores for Test 1, click on cell B31. Then type =AVERAGE(B2:B29) and press the return key. (Be sure to include the equal sign and no space before the word AVERAGE.) This tells the program to calculate the average of the scores that are in cells B2 through B29. The value of the mean (71.39) will appear in the B31 cell. To calculate the standard deviation (*SD*) of scores for Test 1, click on cell B32. Then type =STDEVP(B2:B29) and press the return key. The *SD* (14.65) of the scores in cells B2 through B29 will appear in cell B32. Similarly, if you type =AVERAGE(C2:C29) into cell C31 and press return, the mean of Test 2 (71.25) will appear in cell C31. If you type =STDEVP(C2:C29) into cell C32, the *SD* (11.37) will appear in cell C32.

To calculate the correlation coefficient, you need to tell the program the two columns of scores that are involved. Click on cell B33, then type =CORREL(B2:B29, C2:C29). This tells the program to correlate the Test 1 scores in cells B2 through B29 with the Test 2 scores in cells C2 through C29. Press the return key and the correlation, $r = .75$, appears in cell B33.

FIGURE I.14 Computing basic statistics with the Excel spreadsheet.



	A	B	C
1	Names	Test 1	Test 2
2	Anthony	89	94
3	Ashley	75	68
4	Blake	74	72
5	Chad	84	77
6	Donald	56	66
7	Edward	80	68
8	Festina	66	68
9	George	86	73
10	Harriet	68	73
11	Irene	98	86
12	Jesse	65	78
13	Katherine	44	60
14	Lorraine	45	53
15	Marya	61	75
16	Nancy	75	76
17	Oprah	68	54
18	Peter	55	53
19	Quincy	70	68
20	Robert	69	65
21	Sally	60	47
22	Tina	73	74
23	Ula	75	88
24	Veronica	71	73
25	Wallace	43	61
26	William	83	87
27	Xavier	95	83
28	Yvonne	96	85
29	Zena	75	70
30			
31	Mean	71.39	71.25
32	Std Dev	14.65	11.37
33	Correlation	.75	

Enter students' names and scores into these rows and columns

Type =STDEVP (B2:B29) into this cell

Type =AVERAGE(B2:B29) into this cell

Type =CORREL(B2:B29,C2:C29) into this cell

Computational Procedures for Various Reliability Coefficients

FIGURE J.1 **Example of how to compute the Spearman-Brown double length and the Rulon split-halves reliability estimates.**

A. Pupil's item scores and total test scores[a]

	Items on test				**Total score**
Pupils	**1**	**2**	**3**	**4**	**(*X*)**
Alan	1	0	0	0	1
Isaac	1	1	0	0	2
Leslie	0	0	1	1	2
Miriam	0	0	0	0	0
Rebecca	1	1	0	1	3
Robert	1	1	1	1	4

$$M = \frac{\Sigma X}{N} = \frac{12}{6} = 2;\ (SD_x)^2 = \frac{\Sigma(X - M)^2}{N} = \frac{10}{6} = 1.67$$

[a] An item is scored 1 if it is answered correctly; 0 otherwise.

B. Computation for Spearman-Brown formula

	Half-test scores		**Computing correlation between halves**[b]		
	odd items	**even items**	***z*-scores for:**		**Product**
Pupils	**(1 + 3)**	**(2 + 4)**	**odd**	**even**	**($z_o \times z_e$)**
Alan	1	0	0	−1.22	0
Isaac	1	1	0	0	0
Leslie	1	1	0	0	0
Miriam	0	0	−1.72	−1.22	2.10
Rebecca	1	2	0	+1.22	0
Robert	2	2	+1.72	+1.22	2.10
Means	1.00	1.00			
SDs	0.58	0.82	$r_{nn} = \frac{\Sigma Z_0 Z_e}{N} = \frac{4.6}{6} = 0.70$		

Spearman-Brown
double length
reliability estimates $= \dfrac{2r_{nn}}{1 + r_{nn}} = \dfrac{(2)(.70)}{1 + .70} = 0.82$

[b] Other procedures may be used for computing the correlation coefficient (see Figure I.12).

From Appendix J of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FIGURE J.1 (*continued*)

C. Computation for Rulon formula

	Half-test scores[c]		
Pupils	**odd items (1 + 3)**	**even items (2 + 4)**	**Difference between half-test scores**
Alan	1	0	1
Isaac	1	1	0
Leslie	1	1	0
Miriam	0	0	0
Rebecca	1	2	−1
Robert	2	2	0

Variance of differences $= (SD_{\text{diff}})^2 = 0.33$

Variance of total scores $= (SD_x)^2 = 1.67$

Rulon split-halves
reliability estimate

$$= 1 - \frac{(SD_{diff})^2}{(SD_x)^2}$$

$$= 1 - \frac{0.33}{1.67} = 0.80$$

[c] Neither the Spearman-Brown nor the Rulon formula is restricted to an odd-even split. Other splits may be used (see text).

FIGURE J.2 **Example of how to compute the Kuder-Richardson formula 20 (KR20) and the Kuder-Richardson formula 21 (KR21) reliability estimates.**

A. Computing KR20

	Items on test[a]				
Pupils	**1**	**2**	**3**	**4**	**Total score[b](X)**
Alan	1	0	0	0	1
Isaac	1	1	0	0	2
Leslie	0	0	1	1	2
Miriam	0	0	0	0	0
Rebecca	1	1	0	1	3
Robert	1	1	1	1	4
Fraction passing					$M = 2.0$
each item					$(SD_x)^2 = 1.667$
(p-values)	.67	.50	.33	.50	
$(1-p)$	.33	.50	.67	.50	
$p(1-p)$	.222	.250	.222	.250	$\Sigma p(1-p) = 0.944$

$$\text{KR20} = \left[\frac{k}{k-1}\right]\left[1 - \frac{\Sigma p(1-p)}{(SD_x)^2}\right] = \left[\frac{4}{(4-1)}\right]\left[1 - \frac{0.944}{1.667}\right]$$

$$= (1.333)(1 - .566) = (1.333)(.434) = .58$$

[a] An item is scored 1 if it is answered correctly; 0 otherwise.

[b] The mean and variance are computed in Appendix I.

FIGURE J.2 (*continued*)

B. Computing KR21

Pupils	**Total score**[a] (***X***)
Alan	1
Isaac	2
Leslie	2
Miriam	0
Rebecca	3
Robert	4

$M = 2.0$; $(SD_x)^2 = 1.667$

$$\begin{aligned}
\text{KR21} &= \left[\frac{k}{k-1}\right]\left[1-\frac{M(k-M)}{k(SD_x)^2}\right]\\
&= \left[\frac{4}{4-1}\right]\left[1-\frac{2(4-2)}{4(1.667)}\right]\\
&= (1.333)\left[1-\frac{4}{6.668}\right]\\
&= (1.333)\,(1-.600)\\
&= .53
\end{aligned}$$

[a] The mean and variance are computed in Appendix I.

C. Comparing the values of various reliability estimates for the same test[a]

Estimating procedure	**Numerical value**
Spearman-Brown	.82[b]
Rulon	.80[b]
KR20	.58
KR21	.53

[a] See Appendix I. also.

[b] Based on an odd-even split.

FIGURE J.3 Example of how to compute a coefficient alpha reliability estimate for a set of essay questions or judges' ratings.

	Questions or judges				
Persons	**I**	**II**	**III**	**IV**	**Total score (*X*)**
Aaron	4	3	4	4	15
Dorcas	2	5	5	5	17
Katherine	3	5	5	3	16
Kenneth	1	3	1	1	6
Lee	5	5	5	4	19
Peter	4	3	4	4	15
$(SD_i)^2$ values	1.81	1.00	2.00	1.58	$(SD_x)^2 = 16.89$

$\Sigma(SD_i)^2 = 1.81 + 1.00 + 2.00 + 1.58 = 6.39$

$$\alpha = \left[\frac{k}{k-1}\right]\left[1-\frac{\Sigma(SD_i)^2}{(SD_x)^2}\right] = \left[\frac{4}{4-1}\right]\left[1-\frac{6.39}{16.89}\right] = (1.33)\,(1-.38) = (1.33)\,(.62) = .82$$

FIGURE J.4 Example of computing the general Spearman-Brown reliability estimate.

A. Formula

$$r_{nn} = \frac{nr_{11}}{1 + (n-1)r_{11}}$$

B. Example

Q. A teacher has a 10-item test with reliability coefficient equal to 0.40. What would be the reliability if the teacher added 15 new items similar to those currently on the test?

A. Here $r_{11} = 0.40$ and $n = \frac{25}{10} = 2.5$. (The new test would be 25 items long, hence, 2.5 times as long as the original test.) Thus, the new test reliability is:

$$r_{nn} = \frac{(2.5)\,(0.40)}{1 + (2.5 - 1)\,(0.40)}$$
$$= \frac{1.00}{1 + 0.6} = \frac{1.0}{1.6} = .625$$

C. Results of applying the formula to various values of r_{11} and n

	Number of times original test is lengthened (*n*)				
Original reliability	**2**	**3**	**4**	**5**	**6**
.10	.18	.25	.31	.36	.40
.20	.33	.43	.50	.56	.60
.30	.46	.56	.63	.68	.72
.40	.57	.67	.73	.77	.80
.50	.67	.75	.80	.83	.86
.60	.75	.82	.86	.88	.90
.70	.82	.88	.90	.92	.93
.80	.89	.92	.94	.95	.96
.90	.95	.96	.97	.98	.98

FIGURE J.5 Example of how to compute percentage agreement and the kappa coefficient. The kappa coefficient adjusts the percent agreement for chance agreement that is not related to the assessment procedure.

A. General layout of the data

		Results from Test 1		
		Mastery	**Nonmastery**	**Marginal totals**
Results from Test 2	**Mastery**	a	b	a + b
	Nonmastery	c	d	c + d
	Marginal totals	a + c	b + d	N = a + b + c + d

B. Formulas

$P_A =$ total percentage agreement in figure

$$= \frac{a}{N} + \frac{d}{N} = \frac{a + d}{N}$$

$P_c =$ percent agreement expected because of the composition of the group

$$= \left(\frac{a + b}{N} \times \frac{a + c}{N}\right) + \left(\frac{c + d}{N} \times \frac{b + d}{N}\right)$$

$$k = \frac{P_A - P_C}{1 - P_C}$$

FIGURE J.5 (*continued*)

C. Numerical example

		Results from Test 1		
		Mastery	**Nonmastery**	**Marginal totals**
Results from Test 2	**Mastery**	11	4	15
	Nonmastery	1	9	10
	Marginal totals	12	13	25

$$P_A = \frac{11}{25} + \frac{9}{25} = \frac{11 + 9}{25} = \frac{20}{25} = 0.80$$

$$P_C = \left(\frac{15}{25} \times \frac{12}{25}\right) + \left(\frac{10}{25} \times \frac{13}{25}\right) = \frac{180}{625} + \frac{130}{625} = \frac{310}{625} = 0.50$$

$$\kappa = \frac{0.8 - 0.50}{1 - 0.50} = \frac{0.30}{0.50} = 0.60$$

A Limited List of Published Tests

FIGURE K.1 **Selected published tests.**

Title	Age/Grade level	Publisher[1]	Review[2]
Multilevel survey achievement batteries (group)			
• Iowa Tests of Basic Skills, Forms A, B, C	K–9	RP	**14**:159, **17**:93
• Iowa Tests of Educational Development, Forms A, B, C	Gr. 9–12	RP	**14**:160, **16**:116
• Metropolitan Achievement Tests, 8th ed.	Gr. 1–12	PA	**12**:232, **16**:146
• Stanford Achievement Test, 10th ed.	Gr. 2–12	PA	**13**:292, **16**:232
• TerraNova, 3rd ed.	K–12	CTBMH	**13**:40, **16**:245
• TerraNova, Comprehensive Tests of Basic Skills, 5th ed.	K–12	CTBMH	**11**:81, **14**:383
Multilevel survey achievement batteries (individual)			
• Peabody Individual Achievement Test–R	K–12	PA	**11**:280, **14**:279
• Kaufman Test of Educational Achievement–2nd ed.	1–12	PA	**14**:191, **16**:124
• Wide Range Achievement Test–3	K–12	PAR	**12**:414, **16**:272
Multilevel criterion-referenced achievement tests			
• Degrees of Reading Power–Revised	1–12	QA	**14**:111, **16**:69
• Key Math–3	K–9	PA	**11**:191, **14**:194
Reading survey tests			
• Gates–MacGinite Reading Tests, 4th ed.	K–12	RP	**11**:146, **16**:94
Reading diagnostic tests			
• Stanford Diagnostic Reading Test, 4th ed.	Gr. 1–12	PA	**9**:1178, **13**:294
• Woodcock-Johnson III Diagnostic Reading Battery	K–Adult	RP	**14**:422, **17**:201
Adaptive behavior inventories			
• Vineland Adaptive Behavior Scales–II	0–90 yrs.	PA	**10**:381
Individual general ability/scholastic aptitude tests			
• Bayley Scales of Infant Development, 3rd ed.	1–42 mo.	PA	**13**:29, **17**:17
• Draw A Person: A Quantitative Scoring System	5–17 yrs.	PA	**11**:114, **17**:59
• Kaufman Assessment Battery for Children, 2nd ed.	2–12 yrs.	PA	**9**:562, **16**:123
• Peabody Picture Vocabulary Test–III	2–Adult	PA	**9**:926, **14**:280
• Stanford-Binet Intelligence Scale, 5th ed.	2–Adult	RP	**10**:342, **16**:233
• Wechsler Adult Intelligence Scale, 3rd ed.	16–90 yrs.	PA	**9**:1350, **14**:415
• Wechsler Intelligence Scale for Children, 4th ed.	6–16 yrs.	PA	**12**:412, **16**:262, **17**:197
• Wechsler Preschool and Primary Scale of Intelligence, 3rd ed.	3–7 yrs.	PA	**11**:466, **16**:267

From Appendix K of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

FIGURE K.1 (*continued*)

Title	Age/Grade level	Publisher[1]	Review[2]
Group-administered tests of scholastic aptitude			
• ACT Assessment	Gr. 10–12	ACT	**12**:139
• Closed High School Placement Test	Gr. 8	STS	**8**:26, **14**:80
• Cognitive Abilities Test–Form 6	K–12	RP	**13**:71, **16**:55
• Otis-Lennon School Ability Test, 8th ed.	K–12	PA	**11**:274
• College Board SAT Reasoning Test	Gr. 11-12	ETS	**9**:244
Multiple aptitude batteries			
• Differential Aptitude Test, 5th ed.	Gr. 7–12	PA	**12**:118
Vocational interest inventories			
• Hall Occupational Orientation Inventories	Gr. 3–Adult	STS	**12**:175, **16**:100
• Jackson Vocational Interest Survey–Revised	Gr. 9–Adult	SAS	**14**:187, **15**:129
• Self-Directed Search, online	Gr. 7–Adult	PAR	**14**:345
• Strong Interest Inventory, 4th ed.	16 yrs.–Adult	CPP	**12**:374, **15**:248.

[1] See Appendix L for names and addresses of publishers.

[2] The boldface number is the number of the *Mental Measurements Yearbook* volume; the number after the colon is the entry number. Reviews of previous editions are listed in some cases.

List of Test Publishers and Their Websites

See the current *Mental Measurements Yearbook,* or *MMY,* the Buros Institute of Mental Measurements (http://www.unl.edu/buros), or the Association of Test Publishers (http://testpublishers.org) for additional names and addresses.

American College Testing
Program (ACT)
2201 N. Dodge Street
PO Box 168
Iowa City, IA 52243
http://www.act.org

American Council on
Education (ACE)
Suite 800
1 Dupont Circle
Washington, DC 20036
http://www.acenet.edu

Center for Applied
Linguistics (CAL)
4646 40th Street NW
Washington, DC 20016
http://www.cal.org

The College Board (CEEB)
45 Columbus Avenue
New York, NY 10023-6992
http://www.collegeboard.org

CPP (formerly Consulting
Psychologist Press)
1055 Joaquin Road, Suite 200
Mountain View, CA 94043
http://www.cpp-db.com

CTB/McGraw-Hill (CTBMH)
Publishers Test Service
20 Ryan Ranch Road
Monterey, CA 93940-5703
http://www.ctb.com

Educational and Industrial Testing
Service (EDITS Online)
http://www.edits.net

Educational Records Bureau, Inc.
(ERB)
220 East 42nd Street
New York, NY 10017
http://www.erbtest.org

Educational Testing Service
(ETS)
Rosedale Road
PO Box 6736
Princeton, NJ 08541-6736
http://www.ets.org

Institute for Personality and
Ability Testing (IPAT)
PO Box 1188
Champaign, IL 61824-1188
http://www.ipat.com

Pearson Assessment (PA)
19500 Bulverde Boulevard
San Antonio, TX 78259-3701
http://www.PearsonAssess.com

PRO-ED (PE)
8700 Shoal Creek Boulevard
Austin, TX 78757-6897
http://www.proedinc.com

Questar Assessment, Inc.
4 Hardscrabble Heights
PO Box 382
Brewster, NY 10509-0382
http://www.questarai.com

Riverside Publishing Co. (RP)
425 Spring Lake Drive
Itasca, IL 60143-2079
http://www.riverpub.com

Scholastic Testing Service,
Inc. (STS)
480 Meyer Road
Bensenville, IL 60106
http://www.ststesting.com

Sigma Assessment Systems (SAS)
PO Box 610984
Port Huron, MI 48061-0984
http://www.jvis.com

Slosson Educational Publishers, Inc.
538 Buffalo Road
PO Box 280
East Aurora, NY 14052-0280
http://www.slosson.com

Western Psychological Services
(WPS)
12031 Wilshire Boulevard
Los Angeles, CA 90025-1251
http://portal.wpspublish.com/

From Appendix L of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

Answers to Even-Numbered Exercises

Chapter 1

2. a. F
   b. F
   c. F
   d. F
   e. F
   f. F
4. a. Placement decision
   b. Classification decision
   c. Placement decision
   d. Certification decision

Chapter 2

2. a. Mastery learning target
   b. Developmental learning target
   c. Mastery learning target
4. a. Psychomotor, because it requires perception and judgment of color (some cognitive—need to know how to use the remote, the on-screen programming, and so on)
   b. Cognitive, because the main requirement is understanding of parliamentary procedures (some affective—need to use some interpersonal skills to conduct the meeting successfully)
   c. Affective, because group maintenance requires interpersonal skills (some cognitive—operating without working on the science would not contribute to group maintenance)
   d. Psychomotor, because eye-hand coordination and skill at throwing is the primary target (some cognitive—need to understand what a foul line is)

Chapter 3

2. a. Reliability evidence
   b. Reliability evidence
   c. Reliability evidence
   d. Relationship to other variables (external structure evidence), in this case science course-taking
   e. Content representativeness and relevance (content evidence)
4. a. Content representativeness and relevance (content evidence)
   b. Relationships of assessment results to the results of other variables (external structure evidence)
   c. Reliability over assessors (reliability evidence)
   d. Content representativeness and relevance (content evidence) and types of thinking skills required (substantive evidence)
   e. Content representativeness and relevance (content evidence); secondarily substantive evidence

Chapter 4

2. No, the teacher's claim is not justifiable. The split-halves reliability coefficient, using the Spearman-Brown double length formula, would be 0.57. This is too low a reliability coefficient to expect consistent performance.

$$\frac{2 \times .40}{1 + .40} = \frac{.80}{1.40} = .57$$

4. a. Harry's science grade equivalent is probably between 7.6 and 8.0.
   b. Harry's math and science performance still do not differ. The interval is 7.2–7.6 for math and 7.6–8.0 for science. These intervals still overlap, at one point (7.6).
   c. Jane's (6.8–7.2) and Sally's (8.0–8.4) performance still differ.

Chapter 5

2. a. Mr. Smith scenario—violation of professional responsibility

From Appendix M of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

Sound—testing at the end of a unit of instruction
Unsound—not planning assessment to match classroom learning targets; "cramming" in additional learning; counting in official assessment points for material not related to learning targets for which the students were responsible

b. Ms. Williams scenario—violation of professional responsibility
Sound—testing at the end of a unit of instruction
Unsound—not reviewing the test or key for quality; unquestioning reliance on the "authority" of a publisher; unwillingness to discuss assessment with student and parent

4. a. Ms. Appleton scenario—violation of professional responsibility
Sound—using appropriate accommodations
Unsound—changing students' answers

b. Mr. Pennel scenario—violation of professional responsibility
Sound—using essays and performance assessment (assuming they are used for assessing appropriate learning targets)
Unsound—not matching scoring schemes (whether rubrics, checklists, rating scales, or point schemes) to the learning targets means student performance on the essays or performance assessments may not be interpreted appropriately as indicators of achievement

c. Ms. Dingle scenario—violation of professional responsibility
Sound—being willing to adjust borderline grades
Unsound—not using additional *achievement* information to make the adjustment; using her perceptions/opinions to make the adjustment (causing a "halo effect"); not having a sound rationale to give the students about their grades; making comments about the students that could be perceived as personal rather than as about their achievement; and, if the scenario is read to imply that the students indeed are at the same achievement level, giving two different grades for the same achievement level

Chapter 6

2. Answers will vary. Good answers will categorize and evaluate materials accurately and draw appropriate conclusions.
4. Answers will vary. Good answers will have coherent plans that support sound learning targets with appropriate formative and summative assessment information.

Chapter 7

2. a. Prerequisite knowledge and skills deficits
   b. Identifying student errors
   c. Identifying student errors
   d. Prerequisite knowledge and skills deficits
4. a. Evaluative feedback, judgmental tone
   b. Descriptive feedback, specifying improvement
   c. Descriptive feedback, first-person response
   d. Evaluative feedback, external reward

Chapter 8

2. a. Have the blank toward the end of the sentence.
   b. Be written as a question (with only one answer).
   c. Have only one or two blanks. Have the blank toward the end of the sentence.
   d. Be written as a question (with only one answer).
   e. Be written as a question (with only one answer). Have directions for amount of precision required.
   f. Population of what? Without that, it is impossible to know whether the question assesses an important aspect of the learning targets.
4. a. Avoid verbal clues.
   b. Assess important ideas (not trivia or common sense). Be definitely true or definitely false.
   c. Assess important ideas (not trivia or common sense).
   d. State the source of the opinion, if your item presents an opinion.
   e. Focus on only one important idea or on one relationship between ideas.

Chapter 9

2. Answers will vary. Evaluations and revisions of items should be consistent with the Checklist for Reviewing the Quality of Multiple-Choice Items.
4. Flaws include: Premises and responses are not homogeneous. All responses are not plausible for each premise. Longer statements go in premises, not responses. Directions should clearly state the basis for matching. Avoid "perfect matching." Explanations should be logical. Revisions should address the flaws.

Chapter 10

2. Flaws are listed below. Revisions should address these flaws.
   a. Match the assessment plan (which probably required higher-order thinking regarding students' understanding of prejudice). Require students to apply knowledge to a new situation. Require the students to demonstrate more than recall. Make clear length, purpose, amount of time, and evaluation criteria.
   b. Define a task with specific directions (this is too broad—one appropriate answer might be, "It's horrible!"). Word the question in a way that leads all students to interpret the item as intended. Make clear length, purpose, amount of time, and evaluation criteria.

4. a. Regarding the maximum marks (points) for each question, students should realize they have trouble allocating 50 points for these four questions, which probably can only support 10 or 15 points.
   b. Because there are no rubrics or point schemes associated with the maximum marks, there should be disagreement on how to score Jane's responses. Most likely, there will be more disagreement on the items that have more points. Discussion may point to the fact that there are no descriptions of performance required for each point level. Discussion may also highlight the difference between points for varying degrees of correctness of short-answer questions (#1, #2, #4) and the points for varying degrees of quality on the paragraph (#3). Discussion may also note that the essay question (#3) does not follow the checklist for evaluating the quality of essays; lack of definition of the task contributes to its being difficult to score.

Chapter 11

2. Answers will vary. There should be 17 well-designed tasks and scoring schemes.
4. Answers will vary. There should be 13 well-designed tasks and scoring schemes.

Chapter 12

2. Answers will vary. Analyses of the tasks should be accurate and thoughtful.
4. Answers will vary. Self- and peer-evaluations should use the checklist for judging the quality of rubrics and rating scales.
6. Answers will vary. The portfolio design should follow the six-step procedure suggested in the chapter.

Chapter 13

2. a. Perfect positive discrimination—all upper-group students got the item right, and no lower-group students did.
   b. Positive discrimination—50% more of the upper-group students than lower-group students got the item right.
   c. No discrimination—the same proportion of upper-group students and lower-group students got the item right.
   d. Negative discrimination—50% more of the lower-group students than upper-group students got the item right.
   e. Perfect negative discrimination—all lower-group students got the item right, and no upper-group students did.
3. See chart below.
4. a. Item 2 (and Item 3 as it stands, although it wouldn't discriminate negatively if it were keyed properly)
   b. Item 2
   c. Item 3
   d. Items 2 and 4 (and 3 as miskeyed)
   e. Item 1
   f. Item 2
   g. 45%

Chapter 14

2. Answers will vary.
4. a. Note that the fixed-percentage and total-points criterion-referenced methods are exact. The norm-reference rankings are also exact, although the letter grades arising from them are a matter of judgment. Self-referenced grades are also a matter of judgment.

		Options							
Item number	*Groups*	*A*	*B*	*C*	*D*	*Faulty distractors*	*Miskeying*	*Ambiguous*	*Guessing*
1.	Upper	0	2	*9	0	no	no	no	no
	Middle			*5					
	Lower	1	2	*4	4		p = .60	D = .45	
2.	Upper	2	*7	0	2	Possibly A	no	no	no
	Middle		*4						
	Lower	0	*9	1	1		p = .67	D = −.18	
3.	Upper	9	*1	1	0	no	A	no	no
	Middle		*1						
	Lower	6	*2	2	1		p = .13	D = −.09	
4.	Upper	*5	5	0	1	no	no	B	no
	Middle	*8							
	Lower	*3	3	3	2		p = .53	D = .18	
5.	Upper	3	2	3	*3	no	no	no	Yes
	Middle				*4				
	Lower	3	2	3	*3		p = .33	D = .00	

b.

Pupil	*Self-referencing*	*CR, fixed-percentage*	*CR, total points*	*NR, SS score method*
A	A	80%, B	78% of total points, C	Rank 3, B
B	B	72%, C	71% of total points, C	Rank 4, C
C	A	93%, A	87% of total points, B	Rank 1, A
D	A	90%, A	87% of total points, B	Rank 2, B
E	F	31%, F	28% of total points, F	Rank 10, F
F	D	58%, F	54% of total points, F	Rank 8, D
G	C	54%, F	59% of total points, F	Rank 9, F
H	F	59%, F	60% of total points, D	Rank 7, D
I	F	69%, D	66% of total points, D	Rank 6, C
J	F	68%, D	66% of total points, D	Rank 5, C

c. Most disagreement should be with self-referenced grading, and the next most disagreement with norm-referenced grading. Criterion-referenced grades should agree. Note that some might have given Pupil D a B for the fixed-percentage method, because his average is 89.6667. Typically one would round to the nearest percent (90, or an A on the scale we're using).

d. Self-referenced grading requires the most subjective judgment. Norm-referenced ranking is mathematical, but how many of each grade to give is a judgment call. Criterion-referenced methods are objective once the individual assessments' scores and weights have been set.

Chapter 15

2. Answers will vary.
4. Answers will vary. Analyses should be complete, accurate, logical, and well-supported.

Chapter 16

2. a. Linear standard score (z- or SS-score)
   b. Grade-equivalent score
   c. Developmental or growth scale score (expanded scale score or grade equivalent)
   d. Stanine

4.

Percentile rank	*Stanine*	z_n	*DIQ (SD = 15)*	*T-score*
99.9	9	+3.00	145	80
98	9	+2.00	130	70
84	7	+1.00	115	60
50	5	0.00	100	50
16	3	−1.00	85	40
2	1	−2.00	70	30
0.1	1	−3.00	55	20

Chapter 17

2. These professional organizations have Websites and journals that may lead you to test information. Also, contacting the organization may lead you to experts whom you might contact.
4. a. The test's technical manual
   b. The test's technical manual
   c. Test reviews, for example in the *Mental Measurements Yearbook*
   d. Professional journals (accessed by searching ERIC, PsycINFO, or other databases using the test's name)

Chapter 18

2. a. *OLSAT*
   b. *DAT*
   c. Aptitude test for a specific subject
   d. *KABC-II*
   e. *ACT*
4. a. Attitude
   b. Interest
   c. Attitude
   d. Values
   e. Values

References

Achieve. (2004). *Measuring up 2004: A report on language arts literacy and mathematics standards and assessments for New Jersey*. Washington, DC: Author.

Airasian, P. W. (2001). *Classroom assessment* (4th ed.). New York: McGraw-Hill.

Albertson, B. (1998). *Creating effective writing prompts*. Newark: Delaware Reading and Writing Project, Delaware Center for Teacher Education, University of Delaware.

Alexander, P. A. (1992). Domain knowledge: Evaluating themes and emerging concerns. *Educational Psychologist, 27*, 33–51.

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 185–202). New York: Guilford Press.

Allen, R., Bettis, N., Kurfman, D., MacDonald, W., Mullis, I. V. S., & Salter, C. (1990). *The geography learning of high school seniors*. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

American Association on Intellectual and Developmental Disabilities. (2002). *Definition of intellectual disability.* Retrieved October 12, 2009, from http://www.aaidd.org/content_96.cfm?navID=20

American Educational Research Association. (2002). *Ethical standards of the American Educational Research Association*. Retrieved from http://www.aera.net

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Federation of Teachers et al. (1990). *Standards for Teacher Competence in Educational Assessment of Students*.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: National Council on Measurement in Education. Available from http://www.unl.edu/buros/bimm/html/subarts.html

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved from http://www.apa.org/ethics/homepage.html

Anastasi, A. (1988). *Psychological testing* (6th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*, 199–210.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete ed.). New York: Longman.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42*, 145–170.

Arter, J. A. (1998, April). *Teaching about performance assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.

Arter, J. A., & Spandel, V. (1992). NCME instructional module: Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*(1), 36–44.

Arter, J. A., & Stiggins, R. J. (1992, April). *Performance assessment in education*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Assessment Reform Group. (2002). *Assessment is for learning: 10 principles.* Downloadable from http://www.assessment-reform-group.org

Association for Assessment in Counseling and Education. (2003). *Responsibilities of Users of Standardized Tests (RUST)* (3rd ed.). Retrieved from http://aac.ncat.edu/Resources/documents/RUST2003%20VII%20Final.pdf

Azwell, T., & Schmar, E. (Eds.). (1995). *Report on report cards: Alternatives to consider*. Portsmouth, NH: Heinemann.

Baglin, R. F. (1981). Does "nationally" normed really mean nationally? *Journal of Educational Measurement, 18*, 92–107.

Baker, E. L. (1992). Issues in policy, assessment, and equity. In *Proceedings of the National Research Symposium on Limited*

From the References of *Educational Assessment of Students*, 6/e. Anthony J. Nitko. Susan M. Brookhart.

English Proficiency Student Issues: Vol. 1 and 2: Focus on Evaluation and Measurement. Washington, DC.

Baker, F. (2001). *The basics of item response theory*. College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. Available from http://edres.org/irt

Barker, K., & Ebel, R. L. (1981). A comparison of difficulty and discrimination values of selected true-false item types. *Contemporary Educational Psychology, 7,* 35–40.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4,* 305–318.

Becker, K. A. (2003). *History of the Stanford-Binet Intelligence Scales: Content and psychometrics*. Retrieved July 28, 2005, from http://www.assess.nelson.com/pdf/sb5-asb1.pdf

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1974). *Fifth edition manual for the Differential Aptitude Tests (Forms S and T)*. New York: Psychological Corporation.

Benson, J. (1989). Structural components of statistical test anxiety in adults: An exploratory model. *Journal of Experimental Psychology, 57,* 247–261.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7–74.

Blank, R. K. (2002). Using surveys of enacted curriculum to evaluate quality of instruction and alignment with standards. *Peabody Journal of Education, 77*(4), 86–121.

Blommers, P. J., & Forsyth, R. A. (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Boston: Houghton Mifflin.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1,* 1–12.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain*. White Plains, NY: Longman.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1984). *Taxonomy of educational objectives book I: Cognitive domain*. Boston: Allyn & Bacon.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429–444). Upper Saddle River, NJ: Merrill/Prentice Hall.

Boothroyd, R. A., McMorris, R. F., & Pruzek, R. (1992, April). *What do teachers know about testing and how did they find out?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bracken, B. A. (2002). *Bracken School Readiness Assessment: Administration manual*. San Antonio, TX: The Psychological Corporation.

Bransford, J. D., & Stein, B. S. (1984). *The IDEAL problem solver*. New York: W. H. Freeman.

Brelend, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill*. (Research Monograph No. 11.). New York: College Entrance Examination Board.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38,* 295–317.

Brookhart, S. M. (1991). Letter: Grading practices and validity. *Educational Measurement: Issues and Practice, 10*(1), 35–36.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30,* 123–142.

Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice, 18*(1), 5–13.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education, 8,* 153-169.

Brookhart, S. M. (2009). *Grading* (2nd ed.). Upper Saddle River, NJ: Prentice Hall/Merrill Education.

Brown, R. S., & E. Coughlin. (2007, November). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Buros, O. K. (Ed.). (1938). *The nineteen thirty-eight mental measurements yearbook*. New Brunswick, NJ: Rutgers University Press.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245–281.

Cabrera, N. L., & Cabrera, G. A. (2008). Counterbalance assessment: The chorizo test. *Phi Delta Kappan, 89*(9), 677–678.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: Praeger.

Carlson, S. B. (1985). *Creative classroom testing: Ten designs for assessment and instruction*. Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64,* 723–733.

Carroll, J. B. (1974). The aptitude-achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), *The aptitude-achievement distinction: Proceedings of the Second CTB/McGraw-Hill Conference on Issues in Educational Measurement*. Monterey, CA: CTB/McGraw-Hill, Inc.

Center for the Study of Testing, Evaluation, and Educational Policy. (1992, October). *The influence of testing on teaching math and science in grades 4–12*. Boston: Boston College.

Champagne, A. B., & Klopfer, L. E. (1980). *Using the ConSAT: A memo to teachers*. (LRDC Reports to Educators RTE/4). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Christiansen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Cizek, G. J. (2001). *Setting performance standards*. Mahwah, NJ: Erlbaum.

Clarridge, P. B., & Whitaker, E. M. (1997). *Rolling the elephant over: How to effect large-scale change in the reporting process*. Portsmouth, NH: Heinemann.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, S. A., & Hyman, J. S. (1991). Can fantasies become facts? *Educational Measurement: Issues and Practice, 10*(1), 20–23.

Cole, N. S., & Hanson, G. R. (1975). Impact of interest inventories on career choice. In E. E. Diamond (Ed.), *Issues of sex bias and sex fairness in career interest measurement*. Washington, DC: Career Education Program, National Institute of Education, Department of Health, Education, and Welfare.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). Upper Saddle River, NJ: Merrill/Prentice Hall.

Cole, N. S., & Nitko, A. J. (1981). Instrumentation and bias: Issues in selecting measures for educational evaluations. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art*. Baltimore: Johns Hopkins University Press.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38,* 369–382.

Collis, K. F. (1991). *Assessment of the learned structure in elementary mathematics and science.* Paper presented at the Assessment in the Mathematical Sciences Conference, Victoria, Australia.

Committee on Education and the Workforce. (2005). *Full history of the ESEA effort: Press releases, summaries, and information related to H.R. 1, the Reauthorization of the Elementary and Secondary Education Act*. Retrieved from http://edworkforce.house.gov/democrats/eseainfo.html

Connolly, A. J. (2007). *KeyMath^{TM}–3 Diagnostic Assessment*. Pearson.

Corby, K. (2002). *Tests and testing information.* East Lansing: Michigan State University, University Library.

Council of Chief State School Officers. (2005). *Alignment analysis*. Washington, DC: Author. Available from http://www.ccsso.org/Projects/alignment_analysis

Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform*. Cambridge: Cambridge University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record, 64,* 672–683.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

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

CTB/McGraw-Hill. (2008). *TerraNova, third edition, technical report*. Monterey, CA: Author.

Culler, R. E., & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72,* 16–20.

Cyert, R. M. (1980). Problem solving and educational policy. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research*. Hillsdale, NJ: Erlbaum.

Darden, A. D. (2000). Thoughtful conversations: Student-teacher writing conferences. *Trade Secrets: Teaching Tips for Elementary, Middle, and High School Teachers, 20*(1), 3–5.

Das, J. P. (2002). A better look at intelligence. *Current Directions in Psychological Science, 11*(1), 28–33.

Davey, B., & Rindone, D. A. (1990, April). *Anatomy of a performance task*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Davis, H. A., & Li, J. (2008). *The relationship between high school students' cognitive appraisals of high stakes tests and their emotion regulation and achievement*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Davis, R. V. (1980). Measuring interests. In D. A. Payne (Ed.), *New directions for testing and measurement: Recent developments in affective measurement* (No. 7). San Francisco: Jossey-Bass.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D.Fla.1979)

Debra P. v. Turlington, 644 F.2d 397, 408 (5th Cir.1981) (Unit B)

Debra P. v. Turlington, 730 F.2d 1405 (11th Cir. 1984)

Donlon, T. F., & Angoff, W. H. (1971). The Scholastic Aptitude Test. In W. H. Angoff (Ed.), *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and achievement tests*. New York: College Entrance Examination Board.

Dorans, N. J. (2002). *The recentering of SAT scales and its effects on score distributions and score interpretations*. Research Report No. 2002-11. New York: College Entrance Examination Board.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Stem type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education, 8,* 187–197.

Dudycha, A. L., & Dudycha, L. W. (1972). Behavioral statistics: An historical perspective. In R. E. Kirk (Ed.), *Statistical issues: A reader for the behavioral sciences*. Monterey, CA: Brooks/Cole.

Dunbar, D. A., Float, B., & Lyman, F. J. (1980, November). *Report card revision steering committee final report*. Mount Lebanon, PA: Mount Lebanon School District.

Dunbar, S., Hoover, H. D., Frisbie, D. A., & Mengeling, M. A. (2008). *Iowa Tests of Basic Skills complete and core batteries, Form C: 2005 norms and score conversions*. Chicago: Riverside.

Dunbar, S., Hoover, H. D., Frisbie, D. A., & Oberley, K. R. (2008). *Interpretive guide for school administrators: Iowa Tests of Basic Skills, Levels 5–14, Forms A, B, and C*. Chicago: Riverside.

Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.

Duran, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 573–587). Upper Saddle River, NJ: Merrill/Prentice Hall.

Duschl, R. A., & Gitomer, D. H. (1991). Epistemological perspectives on conceptual change: Implications for educational practice. *Journal of Research in Science Teaching, 28,* 839–858.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice Hall.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Ebel, R. L. (1974). Shall we get rid of grades? *Measurement in Education, 5*(4), 1–2.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Egawa, K., & Azwell, T. (1995). Telling the story: Narrative reports. In T. Azwell & E. Schmar (Eds.), *Report on report cards: Alternatives to consider*. Portsmouth, NH: Heinemann.

Ennis, R. H. (1985). Goals for a critical thinking curriculum. In A. Costa (Ed.), *Developing minds: A resource book for teaching thinking*. Alexandria, VA: Association for Supervision and Curriculum Development.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, Department of Labor, & Department of the Treasury. (1979). Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee section procedures. *Federal Register, 44* (Publication Number: 11996–12006).

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978, August 25). Uniform guidelines on employee selection procedures. *Federal Register, 43* (Publication Number: 38290–38315).

Ergene, T. (2003). Effective interventions on test anxiety reduction: A meta-analysis. *School Psychology International, 24,* 313–328.

Ericsson, K. A., & Simon, H. A. (1999). *Protocol analysis: Verbal reports as data*. Cambridge: Massachusetts Institute of Technology.

Evaluation Center. (1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS)*. Frankfort: Kentucky Institute for Education Research.

Feister, W. J., & Whitney, D. R. (1968). An interview with D. E. F. Linquist. *Epsilon Bulletin, 42,* 17–28.

Feldt, L. S. (1967). A note on the use of confidence bands to evaluate the reliability of a difference between two scores. *American Educational Research Journal, 4,* 139–145.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Ferguson, R. L. (1970). A model for computer-assisted criterion-referenced measurement. *Education, 81,* 25–31.

Fischer, R. J. (1994). The Americans with Disabilities Act: Implications for measurement. *Educational Measurement: Issues and Practice, 13*(4), 17–26, 37.

Fisher, T. H. (1980). The courts and your minimum competency testing program—A guide to survival. *NCME Measurement in Education, 11*(1), 1–12.

Flanagan, J. C. (1967). Functional education for the seventies. *Phi Delta Kappan, 49,* 27–32.

Flanagan, J. C. (1969). Program for learning in accordance with needs. *Psychology in the Schools, 6,* 133–136.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist, 33,* 671–679.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–235). Hillsdale, NJ: Erlbaum.

Forster, M. , & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (ed.), *Towards coherence between classroom assessment and accountability.* Chicago, IL: University of Chicago Press.

Forsyth, R. A. (1976, March). *Describing what Johnny can do* (Iowa Testing Program, Occasional Paper, No. 17). Iowa City: University of Iowa.

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology, 8,* 389–413.

Frederiksen, N. (1984). Implications of cognitive theory for instruction in problem-solving. *Review of Educational Research, 54,* 363–407.

Freeman, D. J., Kuhs, T. M., Knappen, L. B., & Porter, A. C. (1982). A closer look at standardized tests. *Arithmetic Teacher, 29*(7), 50–54.

Frisbie, D. A. (1992). The multiple true-false format: A status review. *Educational Measurement: Issues and Practice, 11*(4), 21–26.

Frisbie, D. A., & Becker, D. F. (1990). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4,* 67–83.

Frisbie, D. A., & Waltman, K. K. (1992). Developing a personal grading plan. *Educational Measurement: Issues and Practice, 11*(3), 35–42.

Fuchs, L. S., & Fuchs, D. (2007). *Progress monitoring in the context of responsiveness-to-intervention*. Retrieved January 15, 2009, from http://www.studentprogress.org/summer_institute/2007/RTI/RTIManual_2007.pdf

Gagné, R. M. (1962). The acquisition of knowledge. *Psychological Review, 69,* 355–365.

Gagné, R. M. (1970). Instructional variables and learning outcomes. In M. C. Wittrock & F. Wiley (Eds.), *Evaluation of instruction*. New York: Holt, Rinehart & Winston.

Gagné, R. M., & Briggs, L. J. (1979). *Principles of instructional design* (2nd ed.). New York: Holt, Rinehart & Winston.

Gagné, R. M., Briggs, L. J., & Wager, W. W. (1988). *Principles of instructional design* (3rd ed.). New York: Holt, Rinehart & Winston.

Gagné, R. M., Major, J. R., Garstens, H. L., & Paradise, N. E. (1962). Factors in acquiring knowledge of a mathematical task. *Psychological Monographs, 76*(7, whole No. 526).

Gagné, R. M., & Paradise, N. E. (1961). Abilities and learning sets in knowledge acquisition. *Psychological Monographs, 75*(14, whole No. 518).

Geiger, M. A. (1997). An examination of the relationships between answer changing, testwiseness, and examination performance. *Journal of Experimental Education, 66,* 49–60.

GI Forum, et al. v. Texas Education Agency, et al. (January 7, 2000). U.S. District Court (Civil Action SA – 97-CA-1278-EP).

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist, 21,* 99–120.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519–521.

Glaser, R. (1968). Adapting the elementary school curriculum to individual performances. *Proceedings of the 1967 Invitational Conference on Testing Problems,* pp. 3–36. Princeton, NJ: Educational Testing Service.

Glaser, R. (1977). *Adaptive education: Individual diversity and learning*. New York: Holt, Rinehart & Winston.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 625–670). Washington, DC: American Council on Education.

Goldman, B. A., & Mitchell, D. F. (Eds.). (2002). *Directory of unpublished experimental measures, Volume 8*. Washington, DC: American Psychological Association.

Gong, B. (2008, February). *Developing better learning progressions: Some issues and suggestions for research and policy*. Dover, NH: Center for Assessment.

Gong, B., Venezky, R., & Mioduser, D. (1992). Instructional assessments: Level for systemic change in science education classrooms. *Journal of Science Education and Technology, 1,* 157–175.

Good, T. L., & Brophy, J. E. (2002). *Looking in classrooms* (9th ed.). Boston: Allyn & Bacon.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement, 44,* 551–561.

Gronlund, N. E., & Brookhart, S. M. (2009). *Gronlund's writing instructional objectives* (8th ed.). Upper Saddle River, NJ: Pearson.

Gulliksen, H. (1986). Perspective on educational measurement. *Applied Psychological Measurement, 10,* 109–132.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20,* 119–131.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 37–50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 51–78.

Haladyna, T. M, Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309–334.

Haladyna, T. M., Nolen, S. B., & Haas, N. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2–7.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5,* 1–16.

Harrow, A. J. (1972). *A taxonomy of the psychomotor domain: A guide for developing behavioral objectives.* White Plains, NY: Longman.

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. (Eds.) (1936). *The construction and use of achievement examinations: A manual for secondary school teachers*. Boston: Houghton Mifflin.

Hembree, R. (1988). Correlates, causes, effects and treatment of test anxiety. *Review of Educational Research, 58,* 47–77.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, W. E. (1990). Fear of failure as a distinctive personality trait measure of test anxiety. *Journal of Research and Development in Education, 23,* 180–185.

Herndon, E. B. (1980). *Your child and testing*. Washington, DC: National Institute of Education, U.S. Department of Education.

Hess, K. (2007). *Developing and using learning progressions as a schema for measuring progress*. Dover, NH: National Center for Assessment.

Higgins, K. M., Harris, N. A., & Kuehn, L. L. (1994). Placing assessment into the hands of young children: A study of self-generated criteria and self-assessment. *Educational Assessment, 2,* 309–324.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE. *Educational Measurement: Issues and Practice, 3*(4), 8–14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice, 3*(4), 16–18.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993). *Directions for administration: Iowa Tests of Basic Skills, Levels 9–14, Forms K and L*. Chicago: Riverside.

Hough, H. (2005). *Tests and measures in social science: Tests available in compilation volumes*. University of Texas at Arlington, Central Library, Health Sciences. Retrieved from http://libraries.uta.edu/helen/test&meas/testmainframe.htm

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253–264.

Ismail, M. (1994). *Development and validation of a multicomponent diagnostic test of arithmetic word problem solving ability for sixth-grade students in Malaysia*. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Iwanicki, E. F. (1980). A new generation of standardized achievement test batteries: A profile of their major features. *Journal of Educational Measurement, 17,* 155–162.

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton, Alberta: Author, Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators*. Thousand Oaks, CA: Corwin Press. Available from http://www.wmich.edu/evalctr/jc

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs*. Thousand Oaks, CA: Sage. Available from http://www.wmich.edu/evalctr/jc

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards: How to improve evaluations of students*. Thousand Oaks, CA: Corwin Press.

Joint Committee on Testing Practices. (1999). *Rights and responsibilities of test takers: Guidelines and expectations*. Retrieved January 23, 2006, from http://www.apa.org/science/ttrr.html

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education (revised)*. Washington, DC: Science Directorate, American Psychological Association. Available from http://www.apa.org/science/fairtestcode.html

Jones, R. W. (1994). *Performance and alternative assessment techniques: Meeting the challenges of alternative evaluation strategies.* Paper presented at the Second International Conference on Educational Evaluation and Assessment, Pretoria, Republic of South Africa.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Karweit, N. L., & Wasik, B. A. (1992). *A review of the effects of extra-year kindergarten programs and transitional first grades* (CDS Report 41). Baltimore: Center for Research on

Effective Schooling for Disadvantaged Students, Johns Hopkins University.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children, Second Edition*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of KABC-II Assessment*. San Francisco: Jossey-Bass.

Keller, F. S. (1968). Goodbye teacher. *Journal of Applied Behavior Analysis, 1,* 79–89.

Keller, F. S., & Sherman, J. G. (1974). *PSI: The Keller Plan handbook*. Menlo Park, CA: Benjamin-Cummings.

Kelly, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17–24.

Kentucky Department of Education. (1993a). *KIRIS Assessment Portfolio: Mathematics, Grade 4, 1993–1994*. Frankfort: Office of Assessment and Accountability.

Kentucky Department of Education. (1993b). *Portfolios and you*. Frankfort: Office of Assessment and Accountability.

Kentucky Department of Education. (1993c). *Teacher's guide: Kentucky Mathematics Portfolio*. Frankfort: Office of Assessment and Accountability.

Keyser, D. J., & Sweetland, R. C. (Eds.) (2005). *Test critiques: Volume 11*. Austin, TX: PRO-ED.

Khattri, N., Reeve, A. L., & Adamson, R. J. (1997). *Studies of education reform: Assessment of student performance*. Washington, DC: Office of Educational Research and Improvement.

King, K. V., Gardner, D. A., Zucker, S., & Jorgensen, M. A. (2004, July). *The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics* (Pearson Assessment Report). Retrieved January 21, 2009, from http://pearsonassess.com/NR/rdonlyres/D7E62EC6-CC3F-47B6-B1CB-A83F341AD768/0/Distractor_Rationales.pdf

Klopfer, L. E. (1969). *An operational definition of "understand."* Unpublished manuscript, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA.

Klopfer, L. E. (1971). Evaluation of learning in science. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Klotz, M. B. & Canter, A. (2006). *Response to Intervention (RTI): A primer for parents.* Retrieved February 5, 2009, from http://www.nasponline.org/resources/factsheets/rtiprimer.aspx

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: Praeger.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Book 2. Affective domain*. White Plains, NY: Longman.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16–25, 46.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Despriet, L. H. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of State School Officers.

Landau, D., & Lazarsfeld, P. F. (1968). Quetelet, Adolphe. *International encyclopedia of the social sciences, 13,* 247–257.

Lane, S. (1992). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice, 12*(2), 16–23.

Lane, S., Parke, C., & Moskal, B. (1992). *Principles for developing performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23–30.

Langenfeld, T. E., & Crocker, L. M. (1994). The evaluation of validity theory: Public school testing, the courts, and incompatible interpretations. *Educational Assessment, 2,* 149–165.

Lindvall, C. M. (1976). Criteria for stating IPI objectives. In D. T. Gow (Ed.), *Design and development of curricular materials: Instructional design articles* (Vol. 2). Pittsburgh, PA: University Center for International Studies, University of Pittsburgh.

Lindvall, C. M., & Bolvin, J. O. (1967). Programmed instruction in the schools: An application of programming principles in "Individually Prescribed Instruction." In P. Lange (Ed.), *Programmed instruction: 66th yearbook, Part II* (pp. 217–254). Chicago: National Society for the Study of Education.

Lindvall, C. M., & Nitko, A. J. (1975). *Measuring pupil achievement and aptitude* (2nd ed.). New York: Harcourt Brace Jovanovich.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1–16.

Linn, R. L. (1994). Performance assessment: Policy, promises, and technical measurement standards. *Educational Researcher, 23*(4), 4–14.

Linn, R. L., & Baker, E. (1997, Summer). CRESST conceptual model for assessment. *Evaluation Comment, 7*(1), 1–22.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5–21.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5–14.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Lord, F. M. (1953). The relation of test scores to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–549.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maddox, T. (Ed.). (2008). *Tests: A comprehensive reference for assessments in psychology, education, and business* (6th ed.). Austin, TX: PRO-ED.

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *Journal of Abnormal and Social Psychology, 47,* 166–173.

Marshall, J. C. (1967). Composition errors and essay examinations grades reexamined. *American Educational Research Journal, 4,* 375–385.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the Dimensions* *of Learning Model*. Alexandria, VA: Association for Supervision and Curriculum Development.

Matter, K. K. (1989). Putting test scores in perspective: Communicating a complete report card for your schools. In L. M. Rudner, J. C. Conoley, & B. S. Plake (Eds.), *Understanding achievement tests: A guide for school administrators* (pp. 121–129). Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Mayer, R. E., Larkin, J. H., & Kadane, J. B. (1984). A cognitive analysis of mathematical problem-solving ability. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*. Hillsdale, NJ: Erlbaum.

McDonnell, L. (1997). *The politics of state testing: Implementing new student assessments. CSE Technical Report 424*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

McKeachie, W. J., Pollie, D., & Spiesman, J. (1985). Relieving anxiety in classroom examinations. *Journal of Abnormal and Social Psychology, 50*, 93–98.

Mealey, D. L., & Host, T. R. (1992). Coping with test anxiety. *College Teaching, 40*, 147–150.

Mehrens, W. A. (1991). Facts about samples, fantasies about domains. *Educational Measurement: Issues and Practice, 10*(2), 23–25.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14–22.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Upper Saddle River, NJ: Prentice Hall.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Miller, M. D., & Seraphine, A. E. (1993). Can test scores remain authentic when teaching to the test? *Educational Assessment, 1*, 119–129.

Millman, J., Bishop, C. H., & Ebel, R. L. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229–258.

Moss., P. A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., & Herter, R. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), 12-21.

Murphy, L. L., Spies, R. A., & Plake, B. S. (2006) *Tests in print VII.* Lincoln: University of Nebraska Press.

National Center on Educational Outcomes. (2009). *Accommodations for students with disabilities*. Retrieved from http://education.umn.edu/NCEO

National Council on Measurement in Education (NCME). (1995). *Code of professional responsibilities in educational measurement (CPR)*. Retrieved from http://www.unl.edu/buros/bimm/html/article2.html

Naveh-Benjamin, M., McKeachie, W. J., & Lin, Y. G. (1987). Two types of test-anxious students: Support for an information processing model. *Journal of Educational Psychology, 79*, 131–136.

Newman, R. (1997–1998). Parent conferences: A conversation between you and your child's teacher. *Childhood Education, 74*, 100–101.

Nichols, S. L., & Berliner, D. C. (2008). Why has high-stakes testing so easily slipped into contemporary American life? *Phi Delta Kappan, 89*(9), 672–676.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447–474). New York: Macmillan.

Nitko, A. J. (1995). Curriculum-based continuous assessment: A framework for concepts, politics, and procedures. *Assessment in Education: Principles, Policy, and Practice, 2*, 321–337.

Nitko, A. J. (2005a). *Using an MMY review and other test materials to evaluate a test*. Retrieved from http://www.unl.edu/buros/bimm/html/lesson02.html

Nitko, A. J. (2005b). *Using an MMY review to evaluate a test*. Retrieved from http://www.unl.edu/buros/bimm/html/lesson01.html

Nitko, A. J., Al-Sarimi, A., Amedahe, F. K., Wang, S., & Wingert, M. (1998, April). *How well are the Kentucky Academic Expectations matched to the KIRIS assessments, the CTBS, and the CAT?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED420677)

Nitko, A. J., & Hsu, T.-C. (1974). Using domain-referenced tests for student placement, diagnosis, and attainment in a system of adaptive, individualized instruction. In W. Hively (Ed.), *Domain-referenced testing*. Upper Saddle River, NJ: Educational Technology Publications.

Nitko, A. J., & Hsu, T-C. (1987). *Teacher's guide to better classroom testing: A judgmental approach*. Pittsburgh, PA: Institute for Practice and Research in Education, School of Education, University of Pittsburgh.

Niyogi, N. S. (1995). *Capturing the power of classroom assessment* (Focus 28). Princeton, NJ: Educational Testing Service.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Midwest Publications, Critical Thinking Press.

Northwest Regional Educational Laboratory. (1998). *Improving classroom assessment: A toolkit for professional developers* (2nd ed.). Portland, OR: Author.

Parkes, J., & Giron, T. (2006). *Making reliability arguments in classrooms.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Parkes, J., & Stevens, J. J. (2002, April). Could school accountability systems be challenged in court? *Educational Assessment Insider: Testing and Accountability, 1*(1), 1–2.

Pearson. (1992). *Integrated Assessment System: Science Performance Assessment*. San Antonio, TX: Author.

Pearson. (2003). *Otis-Lennon School Ability Test (8th ed.) Levels E/F/G. Directions for administering*. San Antonio, TX: Author.

Pearson. (2007). *Stanford Achievement Test Series (10th ed.): 2007 fall supplemental multilevel norms book*. San Antonio, TX: Author.

Pearson. (2008). *WISC-IV and WIAT-II test scores. Report to parents/guardians*.

Pearson. (2009). *Otis-Lennon School Ability Test (8th ed.) Assessing the Abilities That Relate to Success in School*. Retrieved

April 10, 2009, from http://pearsonassess.com/hai/images/dotCom/olsat8/OLSAT_Brochure.pdf

Pearson Education. (1991). *The Differential Aptitude Tests* (5th ed., Fall norms booklet). San Antonio, TX: Author.

Pearson, K. (1924). Historical note on the origin of the normal curve of errors. *Biometrika, 16,* 402–404.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know*. Washington, DC: National Academy Press.

Perie, M., Marion, S., & Gong, B. (2007). *A Framework for Considering Interim Assessments*. National Center for the Improvement of Educational Assessment. Dover, NH: NCIEA. Available at *http://www.nciea.org*

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18,* 16–25.

Perl, J. (1995). *Improving relationship skills for parent conferences. Teaching Exceptional Children, 28*(1), 29–31.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Phoenix, AZ: Oryx Press.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93–120.

Phillips, S. E. (2005, June). Legal corner: Reconciling IDEA and NCLB. *NCME Newsletter, 13*(2). Retrieved from http://www.ncme.org/pubs/vol13_2_June2005.pdf

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10–12, 39.

Popham, W. J. (1991). Appropriateness of teacher's test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12–15.

Popham, W. J. (2005, March). Wyoming's instructionally supportive NCLB Tests. *NCME Newsletter, 13*(1). Retrieved from http://www.ncme.org/pubs/vol13_1_Mar2005.pdf

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states—One hundredth yearbook of the National Society for the Study of Education, Part II* (pp. 60–80). Chicago: University of Chicago Press.

Power, B. M., & Chandler, K. (1998). *Well-chosen words*. York, ME: Stenhouse.

Purves, A. C. (1971). Evaluation of learning in literature. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning* (pp. 697–766). New York: McGraw-Hill.

Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education, 4,* 319–331.

Reschly, D. J. (1993). Consequences and incentives: Implications for inclusion/exclusion decisions regarding students with disabilities in state and national assessment programs. In J. E. Ysseldyke & M. L. Thurlow (Eds.), *Views on inclusion and testing accommodations for students with disabilities* (pp. 35–46). Minneapolis: National Center on Educational Outcomes, University of Minnesota.

Resnick. L. B. (1989). *Tests as standards of achievement in schools*. Paper presented at the ETS Invitational Conference: The Uses of Standardized Tests in American Education, New York.

Robertson, G. J. (1990). A practical model for test development. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 62–85). New York: Guilford Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (2002). Influences on student cognitions about evaluation. *Assessment in Education, 9,* 81-95.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63,* 201–243.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese, 16,* 170–233.

Rudner, L. M., & Boston, C. (1994). Performance assessment. *The ERIC Review, 3*(1), 2–12.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split halves. *Harvard Educational Review, 9,* 99–103.

Ryan, R. M., Connell, J. P., & Deci, E. L. (1985). A motivational analysis of self-determination and self-regulation in the classroom. In C. Ames and R. Ames (Eds.) *Research on motivation in education: Vol. 2. The classroom milieu* (p.13-51). Orlando, FL: Academic.

Sadler, D. R. (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education, 54,* 60-79.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11,* 1–31.

Salend, S. (2009). Using technology to create and administer accessible tests. *TEACHING Exceptional Children, 41*(3), 40–51.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education* (9th ed.). Boston: Houghton Mifflin.

Sanders, N. M. (1966). *Classroom questions: What kinds?* New York: Harper & Row.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46,* 929–938.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49,* 252–279.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Sattler, J. M. (1992). *Assessment of children: WSC-III and WPPSI-R supplement*. San Diego, CA: Author.

Schmeiser, C. B. (1992). Ethical codes in the professions. *Educational Measurement: Issues and Practice, 11*(4), 5–11.

Schutz, P. A., Distefano, C., Benson, J., & Davis, H. A. (2004). The Emotional Regulation During Test-taking scale. *Anxiety, Stress, and Coping, 17,* 253–269.

Scriven, M. (1967). *The methodology of evaluation*. AERA monograph series on curriculum evaluation (Publication No. 1). Chicago: Rand McNally.

Shalaway, L. (1998). *Learning to teach . . . Not just for beginners*. New York: Scholastic Professional Books.

Shavelson, R. J., & Baxter, G. P. (1991). Performance assessment in science. *Applied Measurement in Education, 4,* 347–362.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51,* 455–498.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*(3), 15–22.

Shepard, L. A. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20*(3), 21–23.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: Praeger.

Shuell, T. J. (1990). Phrases of meaning learning. *Review of Educational Psychology, 60,* 531–548.

Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4,* 181–201.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34*(1), 3–12.

Slavin, R. E. (1988). Cooperative learning and student achievement. *Educational Leadership, 46*(2), 31–33.

Smith, J., & Walker, J. (2002). Using electronic gradebooks. *Principal, 82*(2), 64–65.

Smith, M. L. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and teacher capacity building* (CSE Technical Report 425). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Snow, R. E. (1980). Aptitudes and achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade. Proceedings of the 1979 ETS Invitational Conference. New Directions for Testing and Measurement* (No. 5). San Francisco: Jossey-Bass.

Sparrow, S. S., Cicchetti, D. V., & Balla, D., (2005). *Vineland Adaptive Behavior Scales: Second Edition*. Circle Pines, MN: American Guidance Service.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Spearman, C. E. (1927). *The abilities of man, their nature and measurement*. New York: Macmillan.

Stiggins, R. J., Conklin, N. F., & Associates. (1992). *In teachers' hands: Investigating the practice of classroom assessment*. Albany: SUNY Press.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5–14.

Stiggins, R. J., Rubel, E., & Quellmalz, E. (1986). *Measuring thinking skills in the classroom*. Washington, DC: National Educational Association.

Struyven, K., Dochy, F., & Janssens, S. (2005). Student perceptions about evaluation and assessment in higher education: A review. *Assessment and Evaluation in Higher Education, 30,* 325–341.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265–276.

Subkoviak, M. J. (1980). Decision consistency approaches. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore: Johns Hopkins University Press.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263–267.

Swiderek, B. (1997). Parent conferences. *Journal of Adolescent and Adult Literacy, 40,* 580–581.

Tallmadge, G. K., & Wood, C. T. (1976). *User's guide* (ESEA Title I Evaluation and Reporting System). Mountain View, CA: RMC Research Corporation.

Taylor, C. S. (1998). An investigation of scoring methods for mathematics performance-based assessments. *Educational Assessment, 5,* 195–224.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved January 9, 2009, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thorndike, E. L. (1910). Handwriting. *Teachers College Record, 11,* 1–93.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York: Macmillan.

Tippets, E., & Benson, J. (1989). The effect of item arrangement on test anxiety. *Applied Measurement in Education, 2,* 289–296.

Tittle, C. K. (1989). Validity: Whose construct is it in the teaching and learning context? *Educational Measurement: Issues and Practice, 8*(1), 5–13, 34.

Tittle, C. K., Hecht, D., & Moore, P. (1993). Assessment theory and research for classrooms: From taxonomies to constructing meaning in context. *Educational Measurement: Issues and Practice, 12*(4), 13–19.

Touchstone Applied Science Associates. (1995a). *Degrees of reading power and degrees of word meaning: An overview*. Brewster, NY: Author.

Touchstone Applied Science Associates. (1995b). *DRP catalog*. Brewster, NY: Author.

Turner, J. C., Thorpe, P. K., & Meyer, D. K. (1998). Students' reports of motivation and negative affect: A theoretical and empirical analysis. *Journal of Educational Psychology, 90,* 758–771.

United States Supreme Court. (1971). *Griggs et al., Petitioners v. Duke Power Company* (Publication No. 125, 401 U.S. 424, decided March 8, 1971).

U. S. Department of Education. (2005, May 10). *New flexibility for states raising achievement for students with disabilities*. No Child Left Behind. Retrieved from http://www.ed.gov/policy/elsec/guid/raising/disab-factsheet.pdf

U. S. Department of Education. (n.d.). *Accountability*. Retrieved from http://www.ed.gov/nclb/accountability/index.html

U. S. Department of Health and Human Services (USDHHS). (2005, June). *Protection of Human Subjects* (CFR 45, Part 46). Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

Valencia, S. W., & Place, N. A. (1994). Literacy portfolios for teaching, learning, and accountability: The Bellevue Literacy Assessment Project. In S. W. Valencia, E. H. Hiebert, & P. P. Afferbach (Eds.), *Authentic reading assessment: Practices and possibilities*. Newark, DE: International Reading Association.

Viadero, D. (1995). New assessments have little effect on contract, study finds. *Education Week, 14*(40), 6.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64,* 159–195.

Waltman, K. K., & Frisbie, D. A. (1994). Parents understanding of their children's report cards. *Applied Measurement in Education, 2,* 223–240.

Wang, X., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8,* 211–225.

Waples, D., & Tyler, R. W. (1930). *Research methods and teacher problems*. New York: Macmillan.

Webb, N. L. (1999). *Summary report: Alignment analyses of standards and assessments for four states in science and mathematics* (Council of Chief State School Officers and National Institute for Science Education). Madison: Wisconsin Center for Education Research, University of Wisconsin–Madison.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Whitney, D. R., & Sabers, D. L. (1970, May). *Improving essay examinations III: Use of item analysis* (Technical Bulletin No. 11). Iowa City: University Evaluation and Examination Service, University of Iowa.

Wiggins, G. (1990). *The case for authentic assessment* (EDD-TM-9010). Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation, American Institutes for Research.

Wiliam, D. (2007). Content *then* process: Teacher learning communities in the service of formative assessment. In D. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 183-204). Bloomington, IN: Solution Tree.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 132–154). Chicago, IL: 103rd Yearbook of the National Society for the Study of Education, Volume II.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wiser, B., & Lenke, J. M. (1987, April). *The stability of achievement test norms over time*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Woolfolk, A. (2005). *Educational Psychology* (9th ed.). Boston: Allyn & Bacon.

Wormeli, R. (2006). *Fair isn't always equal: Assessing and grading in the differentiated classroom*. Portland, ME: Stenhouse, and Westerville, OH: National Middle School Association.

Young, M. J., & Zucker, S. (2004). *The standards-referenced interpretive framework: Using assessments for multiple purposes* (Pearson Assessment Report). San Antonio, TX: Pearson. Available from http://pearsonassess.com/NR/rdonlyres/1AD36406-3B2A-491C-A280-D4A47D3121E8/0/InterpretiveFrameworks.pdf

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum Press.

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs (Pearson Technical Report)*. San Antonio, TX: Pearson. Available from http://pearsonassess.com/NR/rdonlyres/E5CD33E6-D234-46F3-885A-9358575372FB/0/CognitiveLabs_Final.pdf

Index

Page references followed by "f" indicate illustrated figures or photographs; followed by "t" indicates a table.