Chapter 6

Planning and Constructing Classroom Tests

Learning is not attained by chance; it must be sought for with ardor and attended to with diligence.

—Abigail Adams

Issues and Themes

Testing children in school is one aspect of good teaching. The classroom teacher and his or her students receive the most benefit when the testing is formative. Formative tests are ongoing and fit into the teaching–learning cycle. They are designed to provide information about the details of learning that underlies the progress of individual learners. This information can be used to tailor the instruction to the individual student's needs. These measures tend to be short quizzes, mini-tests, or even informal assessments.

Tests that follow instructional units of study are summative by design. These **summative assessments** are also part of the teaching–learning cycle and are critically necessary for reporting progress to the parents and for recording the success of each individual.

The central task of teachers in constructing classroom tests is to assure that the measures are valid. One important element of validity involves the

content of the test and the test's coverage of the curriculum. Schools that perform well on state assessment tests have studied the state's learning standards, mapped the educational goals and objectives related to those standards, and have built a local curriculum that teaches those standards for learning. The summative tests developed by classroom teachers measure progress toward achieving the goals on objectives outlined in the curriculum. This is facilitated by the use of a two-dimensional test blueprint. The blueprint documents how many questions will be developed to measure each objective and the level of cognition the child will need to answer those questions successfully.

There are various formats that can be used in designing a test measuring achievement. One division is between the use of test items that ask the student to select the correct answer to a question from an array of possible answers and those items that pose a question and ask the student to supply the correct answer. The former of these include true–false tests, matching tests, and three types of multiple choice tests. In all tests, careful planning, and attention to detail, provides the best measurement product.

Learning Objectives

By reading and studying this chapter, you should acquire the competency to do the following:

- Link instruction with the use of formative evaluations in the classroom.
- Provide the documentation needed to support report card grades and conference with parents.
- Devise a curriculum map for a course or grade level.
- Build a table of specifications for an achievement test.
- Explain the six levels of cognition as identified by Benjamin S. Bloom.
- Describe how to construct high-quality multiple choice items.
- List the major pitfalls in the writing of matching-type items.
- Know the advantages of using well-written true–false items.

TESTING AND TEACHING

Good teachers love to teach. If there is anything about the job of being a teacher that is disliked, it is testing and evaluating. One of the first tasks given to graduate assistants by their mentor professors is grading. Likewise, teacher

aids are frequently asked to read and comment on student work for the teacher. Some high schools hire paraprofessionals to read essays written in English classes. Unfortunately, this general dislike of the evaluation process carries over to the mandated statewide assessments. The vast majority of educators see them as a waste of resources and time that would be better spent on instruction. Yet, in this era of accountability, no educator can afford to ignore educational measurement. As Popham (1999) observed, all the rules changed when the test scores from schools appeared in the local newspapers.

Testing and Instruction

The first step in teaching is the identification of the goals for the instructional process. These goals should be expressed in terms of what each child will be able to do or know as a result of the instructional process. Once these goals have been identified and sequentially ordered, instructional activities can begin.

Good teaching involves ongoing monitoring of the learners and the constant tweaking of the instructional process. During every instructional hour good teachers are continually integrating questions into their instruction; they are always looking for those who don't understand or are confused. This ongoing questioning represents an informal approach to assessment. It is informal in that it is idiosyncratic and designed for that particular moment. Informal data collection related to a child can also occur in conversation with other teachers and aids, through observation of the child, by an analysis of errors the child makes, by a careful reading of the homework the child completes, and by asking questions of the child's parents.

FORMATIVE TESTS

Classroom quizzes and tests serve a purpose that is similar to informal assessments. These measures are referred to as formative. The name formative is used because they are designed to inform instructional practice. The best instructional practices involve the use of frequent quizzes designed to inform both the teacher and the learners of achievement and acquired new skills and of areas of learning difficulty. This provides documentation of the effectiveness of the teaching and/or indicates areas needing reteaching (Ainsworth & Viegut, 2006).

Standardized tests and statewide assessments are designed to measure after instruction has occurred. These measures are used to quantify how much was learned by each child. This type of assessment is a summative

measure, as it is given to summarize what has happened. All high-stakes tests and final examinations in secondary schools and colleges fall into this group.

Application of Formative Measures

The type of measurement that is most useful to the teaching process occurs while learning is still underway. These tests are formative in nature, as they can inform the instructional process. Most classroom tests and all quizzes can be formative. The formative testing process makes it possible for a teacher to determine whether students hold a misconception in common or if they have learned and achieved what was expected. This can be as simple as doing a brief analysis of the items from a classroom quiz. After common learning problems have been identified, it is possible for the classroom teacher to reteach the topic, and the instructional outcome will be improved.

The key concept in **formative evaluations** is that the classroom assessments must support the instructional process. This involves much more than deciding on a child's grade. Formative assessment data make it possible to modify and improve the instructional process while it is occurring. In addition, formative evaluations also provide a quality-control system for what is happening in the classroom (Leahy, Lyon, Thompson, & Wiliam, 2005). Each time a teacher asks a student a question, a datum has been collected. When such data are used correctly the nature of student learning can be understood, and the quality of instruction can be improved.

Thus, the learning cycle includes (a) instruction, (b) formative and informal assessment, (c) feedback for the teacher and feedback for the students, (d) modification of instruction and modification of the behavior of the learner. Each of these four steps uses data generated by the assessment. The information provided by the assessment can be as simple as identifying and rewarding a student's good work. It can also involve helping other students to create a learning scaffolding similar to that of the model student (Good & Brophy, 1995).

Formative assessments also can help the instructor in setting the correct pace for learning. This pacing task needs to be accurately set and monitored. When instruction moves too slowly students become bored and disconnected. When the pace of learning is too fast many students may be left behind.

The cognitive depth that students need to employ to learn new material can also be monitored through the use of formative assessment strategies. By asking the students a few questions before instruction begins, each of which is written to a different cognitive level, and asking the questions again during the instructional process, it is possible for the teacher to ascertain if students are employing the optimal learning strategies (Carroll, 1963).

Student Response Pads

A new approach to formative assessment uses an interactive computer system to give teachers real-time formative assessment data during instruction. This new technology has been called "classroom clickers" (Duncan, 2005). To initiate this system, schools first invest in hardware that resembles remote controls for each child (see Photo 6.1). These **student response pads** provide a way for each student to communicate directly with the teacher's computer. During instruction the teacher can ask or in some way present a question and students are able to supply immediate answers using their response pads. This is much like being able to call on every child in the class at the same time. The pulse of the class can be taken by the teacher at any moment and the result seen on a handheld computer screen. In the wireless environment needed for the student response technology, the data can be outputted onto a wireless PDA. A number of textbook publishers have begun to supply classroom response systems as an incentive for schools to purchase their textbooks and curriculum materials.

Case in Point (6a)

The response pad system has been widely demonstrated by several television quiz shows recently. One example was the original ABC network show *Who Wants to Be a Millionaire*. The show is now syndicated by Disney's Buena Vista to most TV markets in the United States. One of the "lifelines," provided by the game rules, allows the contestant to poll the studio audience for their opinion on the multiple choice questions being asked. A bar graph representing the audience's choices appears on the screen. This polling of the audience is completed by a response pad system.

In the classroom the teacher can periodically provide multiple choice questions for the class to answer. The pattern of answers will demonstrate how many students understand the concept being taught. By writing lesson plans with several such marker questions imbedded each day, the teacher can keep close tabs on the learning that is happening in the classroom. The new generation of clickers does not require a "clean line of sight" to operate and facilitate more give and take in the normal classroom setting (Guy, 2007). The total number of response pad units in the schools and colleges of the United States was estimated to total 8 million in 2008. This reflects the falling costs of the systems. It is now possible to purchase classroom clickers for less than $100 per unit (Cavanaugh, 2006).

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Photo 6.1 Students Use Response Pads

SOURCE: From Eduware Inc. www.eduware.com. Reprinted with permission from Bill Stevens, president of Eduware Inc.

Quizzes and tests can also be administered using this system. Another advantage of the use of response pads for classroom instruction is that there is evidence that their use improves student memory and retention of instructional material (Chan, McDermott, & Roediger, 2006). The multiple choice test questions can be presented on paper, but a better model involves the presentation of the items using a PowerPoint format. The response pads have coded signals to the computer identifying each child. The system is able to download student test data into a Microsoft program that will summarize and analyze and save the data in the teacher's grade book.

Special Needs Children

A second function of the formative evaluation process that occurs within the classroom is the early identification of children at risk for failure. The need for vigilance on the part of primary-grade teachers has never been greater. The fact that all third-graders must take a high-stakes test, which in many states could force them to repeat the year, makes it critical that educators watch for children who are struggling to keep up with their classmates. The early identification of children who are at risk for failure can make it possible to initiate remediation and developmental instruction before the children become lost in the system.

Case in Point (6b)

Computers have recently become significantly easier for children with disabilities to use. President Clinton signed into law the Communications Technology Amendments to the Rehabilitation Act in 1998. That law has a set of mandates requiring all computers provide adaptations for users who have a disability (P.L. 105-220, Title 29 U.S.C. 794d § Sec.508). These adaptations include keyboard and mouse modifications for the orthopedically disabled, varying font sizes for the screen, adjustable sound levels, and speech-recognition software.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Learning Standards

One method for improving a school's average scores on state assessment tests is to map out the state's standards and design local methods to measure and track the progress of children toward achieving those learning standards (Guskey, 2005). In this effort, formative assessments can be linked directly with the learning standards and objectives approved by the state and included in the high-stakes assessment. This is more than a simple paraphrasing of the state's exemplar test items, involving an analysis of what the learning standard requires and what students are able to actually do. This analysis must include the underlying skills implied by the standard. Classroom measures designed to assess student performance against the published standards can be thought of as providing benchmarks for progress toward proficiency (Herman & Baker, 2005).

A number of school districts have begun to employ formative testing to reduce the gap between minority and Anglo-White children. The Fairfax, Virginia, schools have adopted the use of a series of formative mini-tests during the school year. These formative tests monitor progress toward the achievement of the standards for learning that are measured by that state's assessment tests. Fairfax's mini-tests are designed to point out problematic vocabulary and undeveloped specialized skills that are required for the state's testing program (Glod, 2006).

Test Preparation

Another reason for introducing both formative and summative testing in the early grades involves the need to familiarize children with the testing format. The contrived environment that is needed for the administration of

high-stakes assessments can be disconcerting to children unfamiliar with the testing process. Young children are comfortable helping each other and sharing answers, behaviors antithetical to high-stakes testing. Young children also expect that their teachers will provide them with answers and help them when they experience difficulty. The regular use of tests and quizzes by primary-grade teachers can introduce children to the reality of the summative testing that they will face starting in third grade.

Parental Reports

Another function of formative classroom evaluations is to provide the documentation needed to submit a report card grade (see Chapter 10) and to prepare for parent–teacher conferences. Parents need and deserve to see work samples from their children during conferences with teachers. Work samples that were part of formative evaluations during the school term can be organized and presented as part of a folio of each child's work and progress. The simple recitation of grades from a roll book will not work for most parents who want to see exactly what problems their children are encountering and what successes they have had. The classroom response system provides printouts depicting graphic data designed to demonstrate both the absolute and relative progress of individual students. It can pinpoint areas of difficulty as well as areas of success. This type of evidence is difficult to debate and can lead to productive and cooperative meetings with parents.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Case in Point (6c)

Not missing an opportunity, in 2006 the Educational Test Service purchased a test consulting company that specializes in formative assessment systems. The entry of ETS into the formative testing field is resulting in a rapid expansion of the availability of new formative classroom assessment approaches. One new product from ETS was the development and sale of access to a bank of thousands of formative assessment items. These items are designed to allow schools to develop their own practice tests for the state-mandated assessment. In 2007–2008, the ETS division on formative assessments became the central player in a national research study on the implementation of formative assessments (Cech, 2007).

SUMMATIVE EVALUATIONS

Children face summative evaluations of their work a number of times each school year. At the end of each high school year, students sit for the dreaded final examinations in their various courses. Elementary and middle-school teachers provide students with tests after each instructional unit. And the state now has a series of mandated achievement tests that children must face during the spring of at least seven of the school years.

Classroom Achievement Tests

A day rarely passes in the lives of school children without their being tested. By the time a student graduates from high school it is likely he or she has taken 2,000 tests, quizzes, and other written evaluations. Research has shown that teachers at all grade levels do not have a high level of confidence in their ability to develop tests of the highest quality (Stiggins & Bridgeford, 1985). The frequent use of classroom quizzes and tests has been demonstrated to be a motivating factor for most children and may even be a direct cause of learning (Slavin, 1994; Tuckman, 2003).[1] The motivational effect is maximized if the time between taking the test and getting a grade or score report is short.

The use of unannounced tests increases both student stress and parent angst (Guskey, 2002; Partin, 2005). Yet, pop-quizzes produce a small achievement advantage and are a favorite technique for ongoing formative assessments by many teachers. By using this strategy, secondary students are forced to always read and prepare before the class, whereas a quiz may or may not happen. Therefore, the teacher makes his or her class the primary one on the student's homework schedule and places his or her assignments ahead the homework assignments made by all the other teachers. In deciding to employ this method, it should be remembered that few opinions and attitudes are as stable over time as those about teachers. One major dimension in teacher evaluations is the perceived "fairness" of classroom tests and evaluations.

Case in Point (6d)

This is an easy truism to verify. If you think back to your school days, it is easy to remember the best and worst teachers you experienced as a student. One factor of the poor teachers may have been a capricious evaluation and

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Case in Point (6d) (Continued)

grading system. This area, evaluation, was also noted by Michael Scriven (1997) as one of particular concern when students evaluate their instructors.

One new, and highly contentious, direction for student evaluations of teaching can be seen on a number of college campuses today. Each year there are more campuses that have online teacher evaluation systems in place. Centralized online systems can reduce the subtle influence that can be exerted by professors who distribute paper-and-pencil forms for student use in the evaluation of the teaching they experienced. These may have been in answer to the movement that was started by student groups on some campuses to provide an independent, but publicly accessible, faculty evaluation system (e.g., Universities of Utah and Mississippi). It is only a matter of time before such systems appear on high school campuses.

Classroom Test Validity

All teachers can build valid and reliable classroom measurements. (A full discussion of these two dimensions is provided in Chapters 4 and 5.) One core issue in the validity of a classroom test is the match between the content that was taught and the questions on the test. When there is a mismatch between the two, students and their parents may be expected to complain. A mismatch is also a sign that the test does not have content validity. Two tools available to teachers to enhance the content validity of classroom tests are **curriculum maps** and a **table of specifications** (test blueprints).

The careful design of classroom tests can also contribute to improving a school's scores on the state-mandated assessments. This is possible when the classroom tests are each linked to the state's approved standards for learning. That set of endorsed learning standards should be at the heart of each school's curriculum. In other words, public schools have an obligation to develop and teach a curriculum aligned to the state standards for learning. This makes it possible for classroom teachers to develop classroom assessments that are an indication of how well the students are achieving the state-required learning standards. Each school should have a central map of its curriculum and how it links to the state standards.

The Mohonasen Central School District (NY) developed curriculum maps that show the sequence and details of what is to be taught in English, mathematics, science, and social studies for each school year.[2] These maps make it possible to develop a series of formative mini-tests that track all children as they progress toward the goal of being graded "proficient" on the statewide mandated assessment (Glod, 2006). Student performance on the

mini-tests can inform the teacher of the need for individualized instruction designed to enhance the learning for children who have not acquired a necessary piece of the curriculum.

Curriculum Mapping

A curriculum map can be thought of as the school's master plan for learning. It should be a publicly available document that captures the scope and sequence of the learning objectives and activities in each subject and in each grade. It is more than a curriculum guide in that it also specifies the skills that students will develop and presents a timeline for instruction in each curriculum area. The map also specifies major educational activities (e.g., field trips and guest speakers) and specifies what assessments are needed for each grade level and classroom.

Curriculum mapping begins with the state learning standards. A team of teachers and curriculum specialists then translates the standards into educational goals and objectives that specify the skills and content to be learned. All subjects and grade levels need to be mapped, even those for which there are no state learning standards.

To review a school district curriculum map for the Spotsylvania, VA, School District see http://205.174.118.254/cmaps/ and from Litchfield, AZ, see http://www.lesd.k12.az.us/curriculum_site/cmap.htm.

The major advantage of curriculum mapping is in providing guidance for the teachers and in the facilitation of instructional planning. Another advantage is that the maps make it possible for teachers to know exactly what was taught the previous year and what will be expected of the children during the ensuing year. Also, parents can quickly see what is being taught and when that instruction is to happen during the school year. For this reason, posting the curriculum map on the school's Web page is a recommended strategy.

Table of Specifications

The construction of valid classroom measures requires planning the test along two dimensions. The first dimension is central to content validity, and the second is the level of cognition the test items will require. These two axes can be plotted together in a test construction blueprint known as a table of specifications (see Table 6.1).

On the vertical axis of the table are specific content objectives for the weather unit. When a classroom teacher builds a test, this axis should exactly reflect the areas of the curriculum that were taught. This process of creating

Table 6.1 A Hypothetical Table of Specifications for a Science Achievement Test of a Weather Unit for Fourth-Grade Children

	Level of Competence Required to Answer the Questions			
Objectives	*Emphasis =*	*Knowledge (40%, 16 items)*	*Comprehension & Applications (40%, 16 items)*	*Analysis, Synthesis, & Evaluation (20%, 8 items)*
Identify cloud forms and explain their formation Emphasis = 40% (16 items)	# of Items =	7	6	3
Understand weather instruments Emphasis = 10% (4 items)	# of Items =	2	1	1
Read and interpret 2 scales on a thermometers Emphasis = 10% (4 items)	# of Items =	1	2	1
Describe and interpret 3 types of weather fronts Emphasis = 15% (6 items)	# of Items =	3	2	1
Explain the "water cycle" Emphasis = 15% (6 items)	# of Items =	2	3	1
Explain tropical storm system Emphasis = 10% (4 items)	# of Items =	1	2	1
TOTAL TEST ITEMS = 40				

a table of specifications prevents the test from overemphasizing those areas for which it is easy to write questions and underemphasizing those areas where test items are difficult to construct. In organizing the table of specifications, the teacher should assign a weight for each content area. This weight serves as a guide in determining the number of items needed for each part of the test. The weight can take the form of a simple percentage, as is shown on Table 6.1. Each of these content areas for which questions are being written should match the instructional objectives specified on the school's curriculum map for the grade level.

Cognitive Requirements

The second axis (horizontal) of the table of specifications indicates the cognitive level that items on this test should require. The question of the level of cognition implied by any learning objective, or needed to respond to any test question, was clarified over 50 years ago. In 1956, Benjamin S. Bloom and other members of a committee of the American Psychological Association provided a taxonomy, or scale of cognition. This scale defines six levels of thought, which are further divided into 19 subcategories. These levels of cognition can be used to classify the level of thinking required by an instructional objective or to answer a test question.

The lowest level of thinking is that which Bloom's committee referred to as "knowledge," and the most complex was "evaluation" (see Figure 6.1).

It is the lower level of cognition (knowledge) that dominates almost all teacher-made tests. Two doctoral dissertations from the 1990s demonstrate this point. Rosemary Castelli (1994) examined the syllabuses and final examinations from 18 freshman and sophomore liberal arts courses at a suburban community college. The courses included those from the humanities, science, mathematics, and social sciences. She found that a preponderance of syllabuses for the courses included course objectives requiring students to use higher order thinking skills. However, the final examinations for all but one of those college classes were dominated by questions requiring the lowest level of cognition (i.e., knowledge). The exception to the rule was an introduction to philosophy class in which the instructor asked examination questions requiring analysis and synthesis.

Maureen Finley (1995) followed up the Castelli study with a project that examined the upper-level courses, including the Advanced Placement classes, of a suburban high school. Here the curriculum guides for the classes had many learning objectives written requiring higher levels of cognition, but the questions on the midyear examinations did not.

Cognitive Level	*Capabilities*	*Assessment Verbs*
Knowledge	Memorization of facts, dates, lists, and events Ability to recall details of subject matter and the vocabulary of the field	Quote directly, who, what, where, when, list, define
Comprehension	Ability to transfer knowledge to other contexts Capability of grasping meaning and predicting consequences, ability to interpret and translate	Interpret, estimate, discuss, translate, summarize
Application	Problem solving skills and the ability to employ theories and concepts to new tasks	Solve, calculate, examine, demonstrate, apply, differentiate, explain
Analysis	Ability to see latent meanings and patterns in information. Capable of ordering parts to see components and commonalities	Sequence and order, select, compare, argue, connect
Synthesis	Using the components from analysis to create new concepts or conclusions. Ability to hypothesize outcomes and to combine concepts into new ideas.	Deductive and inferential thinking, combinatorial, create, plan, invent, design
Evaluation	Assess theories and evaluate constructs and concepts. Compare competing models and make discriminating judgments.	Summarize, assess, judge, evaluate, defend, convince

Figure 6.1 Bloom's Taxonomy

SOURCE: From *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (1st ed.), by Lorin W. Anderson and David R. Krathwohl. Boston: Allyn & Bacon. Copyright 2001 by Pearson Education. Reprinted by permission of the publisher.

There are a number of reasons why teachers do not write questions requiring students to employ higher order thinking skills. Perhaps the most telling are that such questions are difficult to write and take much longer to grade. Typically, these higher order questions require the use of extended answers and an essay (**constructed response**) format.

Bloom also attempted, but never finished, a taxonomy for educators to use for the classification of the affective domain. That taxonomy was written by David Krathwohl and includes the following broad titles: Receiving, Responding, Valuing, Organizing and Conceptualizing, and, at the most advanced level, Characterizing by Value or Value Concept (Krathwohl, Bloom, & Masia, 1964). Another uncompleted effort by Benjamin Bloom was to define a psycho-motor domain. That partial work includes a starting point of Imitation, followed by Manipulation, Precision, Articulation, and, at the highest level, Naturalization (Atherton, 2005).

Cognitive Taxonomies

Bloom's original model has been supplemented by several recent modifications. A less complex model was developed for the teachers of North Carolina in 1989 (Northwest Regional Educational Laboratory, 1989). This model divides the domain of cognition into five levels. Starting at the lowest level, these are recall, analysis, comparison, inference, and evaluation.

Norm Webb, of the Wisconsin Center for Education Research, developed a cognitive taxonomy designed to help align state standards for student learning with statewide assessments. That taxonomy has four levels: (1) recall and reproduction, (2) skills and concepts, (3) strategic thinking, and (4) extended thinking (Webb, 2007).

A student of Benjamin S. Bloom, Lorin Anderson, led a group of cognitive psychologists in the development of a new model for the original taxonomy by Bloom. Their work provided a two-dimensional model (Anderson & Krathwohl, 2001). This system provided one dimension representing the level of knowledge required for the material that was learned, and the other dimension presented six levels of cognitive process needed to perform the mental task (see Figure 6.2).

In addition to this new cognitive taxonomy, Anderson and Krathwohl also proposed four levels of knowledge that could be learned: factual, conceptual, procedural, and meta-cognitive. Their system then plotted the six levels of cognitive process that may be used to answer test questions with the four levels of knowledge, thereby producing a grand total of 24 distinct levels of cognition (four knowledge levels multiplied by six cognitive processes). Thus, the analysis of any test item can be carried out in two dimensions covering 24 distinctly different item types.

Bloom's Taxonomy	*Anderson's Taxonomy*
Knowledge	Remembering
Comprehension	Understanding
Application	Applying
Analysis	Analyzing
Synthesis	Evaluating
Evaluation	Creating

Figure 6.2 Comparison of Bloom's and Anderson's Taxonomies

TESTING FORMATS

Once the table of specifications is complete, the next step is to write the questions that will be a part of the new instrument. The first choice is whether to employ an open-book or closed-book approach. This choice is very much linked to the educational philosophy of the teacher and to the learning objectives for the course.[3] The obvious advantage to employing an open-book test format is that students are put on notice that the test will require analytical answers and that the memorization of lists of facts is not the best approach for test preparation. Another option is the use of an oral-examination system (Lunz & Bashook, 2007). This method for student assessment is often used in graduate and professional schools; also, at the undergraduate level, oral examinations are often employed in assessments of languages and communication.

Beyond the open-book test format is the take-home examination format. Take-home examinations combine elements of a performance evaluation with a structure similar to a traditional examination. This approach is a good method to employ when the expectations for learning are at the highest level and require many hours of analysis to answer. The emergence of virtual schools and online university courses has reignited research interest in take-home examinations (Rakes, 2005–2006).

Conventional wisdom notwithstanding regarding classroom courses and real schools, there is no research evidence that open-book examinations lower test anxiety, reduce cheating, or improve the quality of learning (Cassady, 2001; Kalish, 1958). Lacking any clear evidence of an advantage for the use of open-book and take-home examinations, most teachers at all levels prefer to employ a closed-book, in-class test format.

Item Format

Once the format decision is made, the teacher is then ready to chose between a "**selected response**" or "constructed response" format, or possibly a combination of both (Popham, 1999). Constructed responses are the answers created by the student to questions such as completion, short answer items, essays, and compositions. The latter two can also be considered to be performance items (see Chapter 9).

Selected Response

These test items require the student to select the best answer from several provided on the test. The three major formats for selected items are

true–false, matching, and various permutations of the multiple choice question. There is a general set of test construction rules that apply to all three forms. First, the number of questions should make it possible for more than 90% of all students to complete the test. Girls tend to work at a slightly slower pace on multiple choice format tests than do boys, and the range of scores shows less variance among girls than it does among boys (Longstaffe & Bradfield, 2005). This may reflect the greater tendency for risk-taking by boys (Ramos, 1996). For this reason, girls tend not to do as well on this type of test. The test length should not present an obstacle to those who take a more reflective approach to answering the questions.

Also, each question should have a single focus and use language that is unambiguous and available to all students. The directions should be clear and succinct and provide insight as to how the test will be scored. Students who are impulsive in their tempo tend to miss the directions and begin taking the test as soon as it is in their hands. All test directions should be printed on the test and also read aloud by the teacher. This assures that everyone is following the correct test directions.

True–False Questions

The first of the three principle formats of selected answer questions is true–false. This well-known item type has distinct advantages over other forms of test questions. These items are incredibly easy to score and provide an easy to administer objective measurement of student achievement. Students can answer true–false items at a faster rate than any other format of question. This makes it possible to ask many more questions in a given time period.

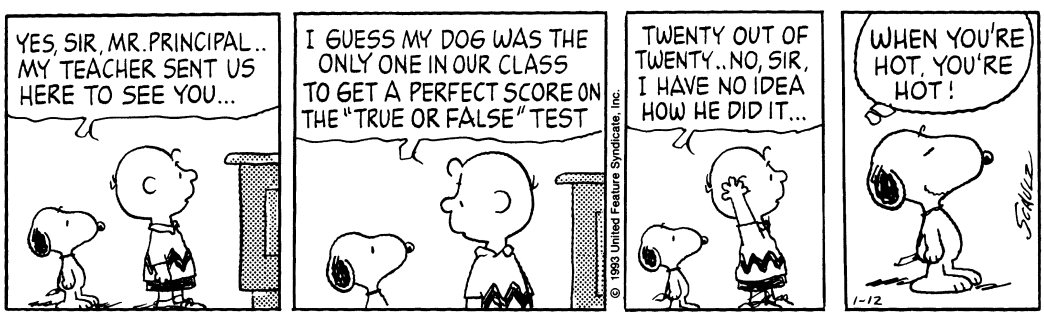

Figure 6.3 Classic Peanuts Cartoon With Charlie Brown and Snoopy

SOURCE: Peanuts: © United Feature Syndicate, Inc.

True–false quiz and test items can serve an instructional role when they are used to motivate students during review sessions. By using a worksheet featuring true–false items, a teacher can identify misconceptions and areas of confusion. Computer-savvy teachers can post true–false questions on their Web page prior to an examination as a way to provide students with a home study tool.

Chief among the disadvantages of using true–false items is the difficulty in writing good questions. Additional problems include the tendency to test at the lowest cognitive level, the ease for student collaboration (cheating), and the large amount of variance that is a function of random guessing. If a student never saw the questions, and only worked with the answer sheet of a true–false format test, the most likely grade would be 50% correct.

A well written true–false item is unequivocally true or false, without any ambiguity. Ambiguity is seen in words such as *sometimes*, *usually*, and *typically*. These words tip off the test taker that the best guess for the answer is "true." Also, in writing such unambiguous items, avoid using exact specification terms or defining words such as *always*, *never*, *not*, *only*, and *must*, which tip off the test taker that the best guess for an answer to the question is "false."

Here are some examples of bad items:

*It **never** rains in Southern California. (F)*

*Water **always** boils at 1,000° C. (F)*

*There are **occasions** when rain falls in the Mojave Desert. (T)*

Here they are again but with better wording:

1. *Southern California receives less annual rainfall than the northern part of the state. (T)*
2. *At standard pressure and temperature (STP) water boils at 1,000° C. (F)*
3. *Las Vegas, NV, receives less annual rainfall than does the Mojave Desert. (F)*

Good true–false test items can be written for any subject area.[4]

1. *When the atmospheric pressure falls, water boils at a temperature greater than 100° C.*
2. *The petitioning employee won in the case (*Hill v. C. A. Parsons *[1972] 1 Ch. 305).*
3. *If $\vec{a}$ is a linear combination of $\vec{b}$ and $\vec{v}$, then $\vec{b}$ is a linear combination of $\vec{v}$ and $\vec{a}$.*

4. *The Socratic educational ideal is based on the view that discovering the truth involves* ***critical thinking****.*
5. *The validity of a test determines the maximum level of reliability the test can have.*
6. *Under International S.T.P. conditions PV = nRT.*
7. *Compound interest accumulates as* $FV = P(1 - r)^n$

One variation of the true–false test has students indicate how certain they are that a question is true or false. Under this system, students indicate T + 3 if they have no doubt if a question is true. The scale extends from F – 3 for being secure in the belief that the item is false, F – 2 for holding a moderate belief in the item being false, and F – 1 for having a high degree of uncertainty about whether the item is false, to T + 1 for being slightly positive the item is true, T + 2 for holding a moderate degree of certainty, and T + 3 for being very positive of a true answer.

Grading of each item is done on a seven-point scale (–3, –2, –1, 0, +1, +2, +3). If the item is skipped and not answered, then it is scored zero. If the item was really true but marked F + 1, the student receives a score of –1 for the question. If the student marked that "true" item T + 2, he or she would be scored +2 for the question, etc.

This complex variation of a true–false test is generally not well received by the majority of students. Students who are risk takers seem to prefer this option. The advantage for the instructor of using this approach is that it results in a less skewed distribution of scores than is typical of true–false tests.

Matching Questions

Another selected format for test items that is frequently used in the middle grades is the **matching question**. This format provides a way to ask questions that requires cognition that is at the comprehension, application, and analysis levels. The questions can be answered quickly, making it possible to cover a wide area of curriculum in a relatively short test. The matching questions can be easy to mark and are less threatening to most students.

The problem with matching items is that the probability of any part of the matching question being correct is influenced by the other sections of the item. If there are five stimulus items to match against seven alternatives, the chances of guessing one correctly is 1:7; however, if the student knows the answer to four of the five stimuli, then the chances of guessing the last stimuli correctly becomes 1:3. These items are difficult to write, as all stimuli and alternatives must be drawn from the same subset of the knowledge

domain. This difficulty may make mixing the areas on the question attractive for the teacher. Matching items are always homogenous in topic coverage. If the various question parts represent two or more distinctly different areas of knowledge, then the item becomes very easy to guess. For example:

Match the items listed in the first column with those in the second column. Not all items in column two are needed to answer this question.

A. Potassium	_____	*Electricity*
B. Lavoisier	_____	*Combustion*
C. Table salt	_____	*NaCl*
D. Ozone	_____	*$NaCO_3$*
E. Franklin	_____	*Magnetism*
	_____	*K*
	_____	*O_3*

In the earlier grades teachers sometimes have children answer matching questions by drawing lines connecting the stimuli and appropriate alternatives. While children like this type of activity, it can present a nightmare of lines and erasures for the teacher to grade. Another common error in constructing this type of test item is to put too many stimuli and alternatives into a matching item. This can create confusion for those children with poor reading and short-term memory skills. The optimal length is five or fewer stimuli and seven or fewer alternatives. The format should have all the longer stems on the same side (the left) and the shorter answers listed on the other. This reduces the need to constantly reread a long list.

It is also important to avoid grammatical clues between stimuli and alternatives. Another example showing common errors of grammatical agreement is shown below. This next question also illustrates another item writing error.

Read each item carefully and match the states on the left column with their official state animal.[5]

STATE	*ANIMALS*
1. The state animals of Pennsylvania are:	*A. Bison*
2. The state animal of Oklahoma is a:	*B. American buffalo*
3. The state animal of Kansas is an:	*C. White-tailed deer & Great Dane*

In this example, a little knowledge of English grammar provides the correct answers. Also, the answers *American buffalo* and *bison* are too similar to be used as alternatives in the same question. A better choice would be to include a state with a dissimilar state animal, such as Alabama, where the state animal is the racking horse.

Multiple Choice Questions (Type A)

Select the best answer from the choices listed below the following question.

Which is true of multiple choice questions (MCQ)?

A. The use of MCQs for testing is widespread throughout education.

B. MCQ items in a test require that the student only recognize the correct answer.

C. MCQ items can be scored in an objective and reliable way.

D. Writing a high-quality test with MCQs is a labor-intensive task.

E. All the above[6]

Multiple choice items like the one above are made of several parts. The question statement is known as the **stem** of the item. The various answers are the **alternatives**, the correct answer is the **keyed** response, and the wrong options are the **distracters**. There are three major formats for MCQ items. The type of item demonstrated above is the standard "type A" item.

The item stem should provide either a complete question or present an incomplete sentence. The former is better for younger children. In reality, most MCQ stems written as incomplete sentences can easily be rewritten as complete questions. The stem should pose only one problem that can be read and clearly understood. The length of the stem should not require students to reread it several times to understand and remember the task before they read the alternatives.

In writing the stem, avoid including negatives. The simple insertion of the word *not* into a question can magnify the item's difficulty and flummox impulsive students, who may never see the word. See the following two possible sample stems from a question on a social studies test about the Civil War. (Asterisk denotes correct answer.)

From the following list of cities, which was a state capital for a Confederate state in 1862?

Or this alternative stem:

Which of the following cities was not a state capital for a Union state in 1862?

*A. Columbia**

B. Sacramento

C. Augusta

D. Springfield

Poor stems leave the student wondering what is expected. The following is an example of such a murky item stem.

Newton's Second Law . . .

A. electricity

B. bodies at rest

C. time

*D. movement**

Here's a better item:

Newton's Second Law explains the relationship between mass and acceleration in terms of what property?

A. weight

*B. force**

C. length

D. light

The alternatives include one correct (keyed) answer and two to four distracters (wrong answers). All the alternatives should be plausible and related to the same topic. All too often a teacher writing a multiple choice item starts with a correct answer then writes a stem for the item. Next, the item author casts about for the distracters. The result is often an odd collection of unrelated nouns that can appear to be the product of free-association.

The best choices for distracters are errors that the students have made previously on homework or misconceptions that came up during class. When the teacher is writing distracters it is not the time to become whimsical or

attempt comedy. The unexpected humor can interfere with the focus and attention of the students to the task. Research has shown that optimal item reliability occurs when there are three options for each item. One option is the correct answer (**keyed answer**) and the other two plausible alternatives (Rodriguez, 2005). However, most published tests provide four or five answer options on each question.

Four other guidelines for writing MCQs include ideas about "all and none," length, sequence, and grammar. Try to always avoid the use of "all of the above" or "none of the above" as an alternative. Both of these alternatives lead to classroom "lawyering," as students make a case after the test is returned that there is really a better answer out there somewhere (none of the above) or that there is something about each alternative that is partially correct (all of the above). Also, the option of "all of the above" is logically incompatible with test directions to "select the one best answer." There are times when there are only three options. The next item is an example.

We are all familiar with the observation that the sun rises in the eastern sky and sets in the west each day. What is the explanation for this phenomenon?

*A. The earth rotates on its axis from west to east.**

B. The earth rotates on its axis from east to west.

C. The earth doesn't move and the sun rotates about it.

In this example, "all or none of the above" would not work as a fourth option.

One way to telegraph the students which of the alternatives is correct is to make it longer than the other options. Testwise, students know that "when all else fails, select the longest answer." This occurs as the question author tends to overwrite the keyed answer to guarantee it is indeed correct. As in the case of matching items, it is important not to alert students to the correct answer by providing grammatical clues. Also the location of the keyed answer should be random. There is a tendency to place the correct answer in position A of the alternatives as it is the first answer that the teacher writes.

In the animal kingdom a frog is classified as **an**_____

*A. amphibian**

B. reptile

C. marsupial

D. fish

The following hypothetical examination is one in which you can get every item correct knowing the rules for making good multiple choice questions.

Please read the following questions and select the best answer to each question.

1. *The purpose of cluss in furmplaling is to remove?*
   - *A. cluss-prags**
   - *B. tremalls*
   - *C. cloughs*
   - *D. plumots*
2. *Trassing is true when?*
   - *A. luso tresses the vom*
   - *B. the viskal flens, if the viskal is conwil or scrtil**
   - *C. the belga frulls*
   - *D. diesless kils easily*
3. *The sigia frequently overfesks the tralsum because of what?*
   - *A. all siglass are mellious*
   - *B. siglas are always votial*
   - *C. the tralsum is usually tarious**
   - *D. no trelsa are feskaole*
4. *The fribbled breg will minter best with an?*
   - *A. derst*
   - *B. marst*
   - *C. sartar*
   - *D. ignu**
5. *Among the reasons for tristal doss are?*
   - *A. when the doss foged the foths tristaled**
   - *B. the kredges roted with the orats*
   - *C. few rekobs accepted in sluth unless foths are present*
   - *D. most of the polats were thronced not tristaled*
6. *Which of the following (is, are) always present when trossalls are being gruven?*
   - *A. rint and vost*
   - *B. vost**
   - *C. shum and vost*
   - *D. vost and plone*

7. *The minter function of the ignu is most effectively carried out in connection with?*
   A. a razama taliq
   B. the thrusting bding
   *C. the fribbled breg**
   D. a frally rnoz

8. *???*
   A.
   B.
   C.
   *D. **

Key to the answers:

1. *The word* cluss *is in the stem and in the answer.*
2. *The conditional "if" gives it away; also it is the longest answer.*
3. *The word* usually *is the giveaway here.*
4. *Here the key is found in the agreement of "an" with "ignu."*
5. *The answer is in the stem.*
6. *The word* vost *is in all other answers.*
7. *The answer in question 4.*
8. *The pattern of the other correct answers is A, B, C, D, A, B, C; ergo, D.*

Each multiple choice test item is one that can be guessed. When a student knows one or two alternatives are wrong, it is possible to improve the chance of being correct when guessing. Yet, every item can be guessed even if the student has no idea which answer is correct. In that unfortunate case, the likelihood that the student will guess the correct answer is a function of the number of answer alternatives. If there are four alternatives, the chance of a correct, albeit blind, guess is 1:4. Should this occur on every item, the unlucky student will likely score near this "**chance level**," or about 25%. If there are five alternatives, the chance of a correct guess is 1:5. Knowing this fact, a number of test publishers have a correction formula used to correct for guessing (Lord & Novick, 1968).

This is done by counting the number of wrong answers on the test and subtracting those that were not attempted. The number remaining is the number attempted but answered wrong. This value is then divided by the number of alternatives. This provides an estimation of the number of items

correctly guessed on the test. This value (the number answered wrong divided by the number of alternatives) is then subtracted from the total number of questions answered correctly. This is then reported as the raw score "corrected for guessing." Because boys are more likely to take blind guesses than are girls, this approach may produce a differential effect by gender (Ben-Shakhar & Sinai, 1991).

All students deserve advanced warning if this scoring method (correction for guessing) is used. When the test is scored using a correction for guessing, the best strategy for test takers to employ is to skip any item where they have no knowledge about any of the alternatives. If one or more of the alternatives can be eliminated, then it will pay the student to make an "educated guess" between the two alternatives that are possible.

For more information, see "Considerations on Point" at www.sagepub.com/wrightstudy

Case in Point (6e)

The publishers of most textbooks provide teachers with online banks or CD-ROMs with test questions as well as printed books of test items. Occasionally these item banks are even organized by instructional objectives and item difficulty. While there are no absolutes, the quality of these items is generally poor. One reason for this is that the authors of the texts rarely write their own item manuals or item banks. A large number of full-time graduate assistants have supplemented their incomes by writing the questions to accompany new books.[7]

The second and third formats for multiple choice items are type K and type R items. Both of these are easily adapted for writing questions requiring higher order thinking skills. These items are extensively used on the qualifying examinations of various medical and other professional licensing boards.

Multiple Choice Questions (Type R)

This format for multiple choice items is sometimes referred to as an extension of the well-known matching question. This format requires that test takers have good reading comprehension skills. The format is used in qualifying examinations in medical disciplines. It also has many applications when there is a large body of high-consensus material that can be used in item development. This includes most science disciplines and a number of areas in the social sciences.

Extended-matching items are multiple choice items organized into sets that use one list of options for all items in the set. A well-constructed extended-matching set includes four components:

1. A theme, or the subject area being assessed by the items.
2. An option list, or a list of all the potential **alternative answers**.
3. A lead-in statement that provides a common task to complete based on the question posed by each item stem.
4. A list of two or more item stems (questions) that can be answered by the options.

Theme: Microbiology
Options:

A. Adenovirus

B. Aspergillus fumigatus

C. Bacillus anthracis

D. Candida albicans

E. Chlamydia psittaci

F. Coccidioides immitis

G. Coronavirus

H. Corynebacterium diphtheriae

I. Coxiella burnetii

J. Coxsackievirus

K. Epstein–Barr virus

L. Haemophilus influenzae

M. Histoplasma capsulatum

N. Mycobacterium tuberculosis

O. Mycoplasma pneumoniae

P. Neisseria gonorrhoeae

Q. Neisseria meningitides

R. Pneumocystis carinii

S. Rhinovirus

T. Streptococcus pneumoniae

U. Streptococcus pyogenes (Group A)

Lead-in:

For each patient who has presented with fever, select the pathogen most likely to have caused his or her illness. Select the one most likely pathogen.

Stems:

1. A 7-year-old girl has a high fever and a sore throat. There is pharyngeal redness, a swollen right tonsil with creamy exudates, and painful right submandibular lymphadenopathy. Throat culture on blood agar yields numerous small hemolytic colonies that are inhibited by bacitracin.

Answer: U

2. For the past week, an 18-year-old man has had fever, sore throat, and malaise with bilaterally enlarged tonsils, tonsillar exudates, diffuse cervical lymphadenopathy, and splenomegaly. There is lymphocytosis with atypical lymphocytes. The patient tests positive for heterophil antibodies.

Answer: K

Multiple Choice Questions (Type K)

Type K multiple choice items allow the test developer to ask questions with more than one correct answer. This format of question was used on early forms of the SAT. In the recent past this item type was used primarily for questions on medical board qualification tests. The use of these items has generally been supplanted by the use of R-type multiple choice items. Students who are asked to answer this type of question must have a good level of reading comprehension and a good memory. The stem that sets up the question must be memorized and then compared with each of the complex list of possible answers.

Type K questions provide a set of directions that are repeated at the top of every page of the test where these items appear. This multiple choice format also has a stem that is expressed as a problem for the test taker to solve. The third part of the question provides the answer alternatives.

DIRECTIONS

For the following questions, *one or more* of the alternatives given are correct. After deciding which alternatives are correct, record your selection on the answer sheet according to the following key:

Mark A if alternatives 1, 2, and 3 only are correct.

Mark B if alternatives 1 and 3 only are correct.

Mark C if alternatives 2 and 4 only are correct.

Mark D if alternative 4 only is correct.

Mark E if all four alternatives are correct.

SUMMARY OF DIRECTIONS

A	B	C	D	E
1, 2, 3 correct	1, 3 correct	2, 4 correct	4 correct	All are correct

STEM

A 27-year-old woman complains of diplopia and difficulty maintaining her balance while walking. These symptoms began suddenly yesterday morning while she was bathing. During the past 2 years she has also had several day-long episodes of blurred vision in one or both eyes. She remembers that one of these episodes also began while she was bathing. In addition to a thorough clinical examination, an appropriate initial diagnostic workup would include the following:

1. cortical evoked response studies
2. basilar arteriogram
3. cerebrospinal fluid protein
4. bone marrow biopsy

Correct answer: 2

Summary

The development and proper use of high-quality tests is a part of good teaching. Ongoing quizzes and the use of classroom clickers are part of formative evaluations, while achievement tests are a form of summative assessment. Teachers can use tests to identify learning problems, individualize instruction, prepare children for future high-stakes tests, and collect data to share with parents.

The best classroom achievement measures are designed following a plan that includes a two-dimensional blueprint. The blueprint provides both

content and cognitive level guidance for the teacher and can result in a valid measure.

Select items, such as multiple choice questions, provide an opportunity to test a wide swath from the curriculum that students have studied. The writing of quality items requires time and practice on the part of the teacher. While students cannot bluff on a multiple choice test, they can guess the correct answer by being test wise and understanding a few simple rules.

Discussion Questions

1. Examine the use of tests in a school with which you are familiar. Do the teachers of this school use tests in a formative way, or are all tests used as summative achievement statements?
2. What points would you make in your presentation if you were asked to provide an inservice lecture on the advantage of formative testing in the schools?
3. Write a multiple choice test item based on this chapter of the textbook that requires test takers to employ Bloom's cognitive level of either "application or analysis" to answer. Then rewrite the same question so that it requires Bloom's level of either "synthesis" or "evaluation" to answer.
4. Write an outline of a presentation you would make for high school students on how to do well on multiple choice tests. Include in your outline issues of wording, guessing, and grammar.
5. Assume that you are asked to explain to your school's parent organization why the purchase of interactive classroom communication systems (clickers) would be a good idea that will benefit the students.

Student Study Site

Educational Assessment on the Web

Log on to the Web-based student study site at www.sagepub.com/wrightstudy for additional Web sources and study resources.

NOTES

1. The level of motivation to do well on a classroom test (low stakes) is related to many factors, including the gender of the children (Eklöf, 2007). Boys tend to have less motivation but a higher level of confidence in their knowledge base than do girls.
2. See www.mohonasen.org/03curriculum/curriculumhome.htm.
3. Some professors of educational research who emphasize application and have little need for students to recall facts use a take-home examination model. This format can reduce student anxiety in complex subject areas while still determining if the student can handle complex tasks.
4. Answers: 1. T, 2. T, 3. F, 4. T, 5. F, 6. T, 7. F
5. Answers: 1. C, 2. B, 3. A
6. Yes, I know it is bad form using "all the above."
7. All of the examination items supplied to faculty teaching this course were written by the text's author.