Improving Measurement and Evaluation

IN HIGHER EDUCATION

BY BENJAMIN E. BANDIOLA

Measurement and evaluation are vital parts of education. State certification programs require that preservice teachers study methods for classroom assessment.[1] Curriculum guidelines produced by professional organizations also support this requirement.[2]

However, liberal-arts majors, many of whom eventually become college professors, do not have to take courses in this area. Justifications for this usually include their subject matter expertise and the fact that they hold advanced degrees in their chosen fields. But good teaching demands more than mastery of content. It requires a variety of other skills, including effective presentation, construction of objectives and syllabi, counseling, and evaluation. Although college-level teachers are not required to take a course in testing and evaluation, they would benefit from learning how to structure their instruction and testing for greater effectiveness.

Picture Removed

While preparing to become an elementary teacher, I took a required course in tests and measurements. I naively expected my college teachers to apply its principles. One professor whom I admired could lecture without notes for the whole period on the history of the Bible. When it was time for his first test, I was well prepared. The test con-

sisted of 40 true-or-false statements. Reading each statement carefully, I found that the first 15 statements appeared to be true—in fact, I could find few questions in the whole test that appeared to be false. This seemed inconsistent with what I had learned in tests and measurements—that there should be a proportionate number of true and false statements, randomly arranged. After analyzing each statement, I changed some of my answers and marked five statements as false.

When the test papers had been collected, the professor asked what we thought of the test. I replied quickly that most of the statements appeared to be true. Looking chagrined, the professor revealed that *all* of the statements were true. He added that he did not believe in putting false statements in a Bible test! I protested that this was not in harmony with what I had learned in tests and measurements class. Nonchalantly, the professor replied, "I have never heard such a thing in my graduate classes."

There is good reason to believe that similar experiences happen regularly in college classes. Since society requires elementary and secondary teachers to achieve competence in tests and measurements, should we expect less of college instructors and professors?

Excellent results have been produced by faculty development programs that include in-service training in measurement and evaluation strategies. The following suggestions will also help professors to polish their measurement skills.[3]

Basic Elements in Education

Every educational program has three interrelated elements: objectives, instructional procedures, and evaluation. Their relationship is illustrated by an equilateral triangle devised by Furst[4] and modified by the author. (See Figure 1.)

As shown by the arrows, the three elements are interdependent. To a great extent, objectives determine instructional procedures and evaluation. Objectives are derived from values that society has attached to education. They develop from the framework of philosophy. For example, What is a good life? How should it be lived? What are the characteristics of an educated person? These philosophical concerns shape teaching objectives. Instructional procedures are based on the psychology of learning, while psychology serves as a laboratory for testing instructional procedures. Evaluation, then, must be developed from the knowledge base of measurement.

Although college-level teachers are not required to take a course in testing and evaluation, they would benefit from learning how to structure their instruction and testing for greater effectiveness.

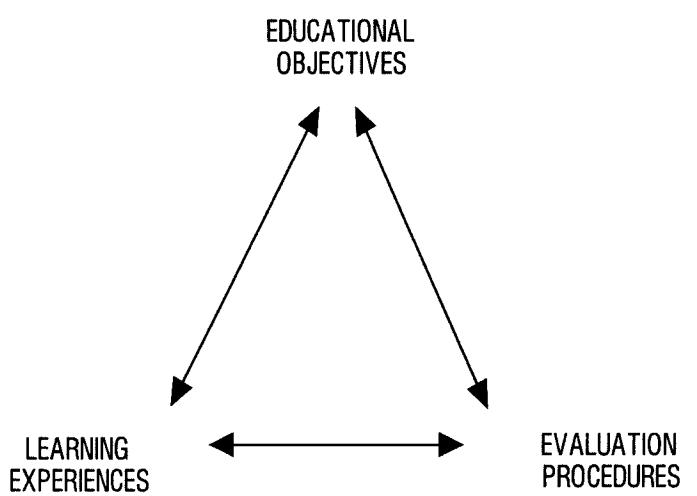
Taxonomy of Educational Objectives

As shown in Figure 1, objectives play a central role in the learning process. However, communicating educational objectives has presented problems, since different authors have interpreted objectives differently and therefore could not agree on how to evaluate them. To resolve this difficulty, the directors of testing services from several midwestern universities gathered at the University of Chicago under the chairmanship of Benjamin Bloom and developed what is now known as "The Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain."[5] This work has become an indispensable reference for curriculum development and test construction. It divides educational objectives into three domains: cognitive, affective, and psychomotor.

The *cognitive domain* refers to intellectual skills and abilities commonly involved in knowing. The *affective domain* is concerned with values, attitudes, interests, and appreciation. These are elements involved in feeling. The *psychomotor domain* deals with motor skills and abilities involved in physical manipulation and activities. These classifications parallel well this philosophical statement by Ellen G. White:

Our ideas of education take too narrow

FIGURE 1
Reciprocal Relationship Among Three Elements



and too low a range. There is need of a broader scope, a higher aim. True education means more than the pursual of a certain course of study. It means more than a preparation for the life that now is. It has to do with the whole being, and with the whole period of existence possible to man. It is the harmonious development of the physical, the mental, and the spiritual powers.[6]

This article emphasizes the cognitive domain. Evaluation of the affective and psychomotor domains presents specialized and technical difficulties. Disciplines that deal with affective and psychomotor objectives have developed specialized methods and instruments for measurement and evaluation. Krathwohl, Bloom, and Masia[7] developed a taxonomy that describes objectives reflecting underlying emotions, feelings, or values. Simpson[8] derived a similar taxonomy for psychomotor behaviors.

According to Bloom's taxonomy, there are six classes of objectives in the cognitive domain: knowledge, comprehension, application, analysis, synthesis, and evaluation. These intellectual skills and abilities are arranged in hierarchical order, based on the complexity of mental processes required for their acquisition. (See Figure 2.) As shown in the diagram, knowledge is the lowest and most common level; evaluation is the highest and least common level.

In higher education, as in the K-12 setting, knowledge (the lowest level) is the most commonly tested, while evaluation (the highest level) is least commonly evaluated. Within this hierarchy, testing for higher-level objectives automatically covers lower-level objectives; however, the reverse is not true. Obviously, then, testing should emphasize higher-level objectives.

Tests are, of course, an important way to measure student achievement. However, they are not the only way that students can demonstrate their knowledge and progress. Because every class has different learning styles and strengths, you should employ a variety of methods for evaluating student progress and computing grades. No quarter/semester grade should be based solely on a final exam, as this puts too much emphasis on a single evaluation that comes too late in the term for students to adjust their level of effort. Students who have learning disabilities or are extremely anxious may not perform up to their true skill level in such cases. Final grades should also factor in multiple student submissions such as portfolios, quizzes, term papers, projects, and short essays. Be sure to respond to each item promptly and in some detail, so that students understand what is expected of them and can prepare for upcoming tests.

Test construction should follow seven distinct steps: (1) planning, (2) writing items that match instructional objectives, (3) assembly, (4) administration, (5) scoring, (6) analyzing results, (7) returning the test and debriefing students.

Tests will doubtless always constitute an important part of course evaluation. Although the principles listed above apply to all kinds of tests, in this article we will focus on ways to construct and improve multiple-choice, objective tests. This does not imply that other types of test should not be used. However, short-answer or essay examinations are somewhat easier to construct.

Steps in Test Construction

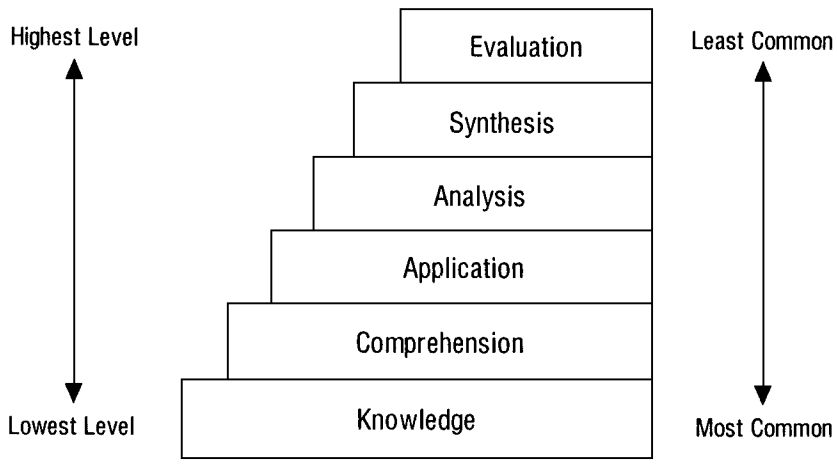
Test construction should follow seven distinct steps: (1) planning, (2) writing items that match instructional objectives, (3) assembly, (4) administration, (5) scoring, (6) analyzing results, (7) returning the test and debriefing students. These steps may require up to 25 percent of the total instructional time, depending on the teacher's background in tests and measurements.

Planning

Constructing good tests is a science and an art. The process follows well-defined steps and allows creativity and self-expression.

Good planning requires a test blueprint or two-way table of specifications. It takes into account both the objectives and the content covered in the course. First, choose objectives based on Bloom's taxonomy. Class topics that received greater emphasis should take up more

FIGURE 2[9]
Taxonomy of Educational Objectives: Cognitive Domain



items in the test, while less-important content should have fewer test questions. Figure 3 shows a blueprint for a test in a tests-and-measurements class. The unit topic was criteria of satisfactory measuring instruments.

According to this blueprint, the test will measure four objectives in Bloom's taxonomy—knowledge, comprehension, application, and evaluation. The test items will deal with three main criteria for measuring instruments—validity, reliability, and accuracy. The test will contain 45 items, with the number of items for each area being determined by the emphasis given to it and the instructional time spent on the topic. Since validity is the most important criterion, the greatest number of items is devoted to it. Other criteria are allocated questions proportionate to the emphasis they received.

Recognizing the appropriate concept from among several options demands a higher level of mental activity than mere recall of information. But it also requires the lower levels of Bloom's taxonomy. One has to *know* something before he or she can understand and use it.

The importance of application was recognized by Ellen G. White:

Every youth should be taught the necessity and the power of application. Upon this, far more than upon genius or talent, does success depend. Without application the most brilliant talents avail little, while with rightly directed effort persons of very ordinary natural abilities have accomplished wonders.[10]

If testing focuses on factual knowledge, then students will spend more time memorizing this type of information. Tests requiring only factual recall breed intellectual mediocrity. Students preconditioned by this type of testing find it difficult to rise to the higher level of thinking required for application.

Evaluation occupies the highest level in Bloom's taxonomy because it requires the most complex mental activity. As the basis of decision-making, it uses knowledge, comprehension, and application.

Writing Items That Match Instructional Objectives

One reason essay examinations are popular with teachers is that they are relatively easy to write. One can construct a valid test by simply converting an instructional objective into an essay-test item. If an instructional objective states that the student will be able to tell the difference between a metaphor and a simile, a valid test item would be "Differentiate between a metaphor and a simile."

Writing valid multiple-choice items requires technical knowledge and skills—and preferably a course in tests and measurements. Multiple-choice items have two parts. The stem specifies the task or problem that the student must perform; the options or alternatives list the responses. Only one choice is the correct answer; the others are distractors or decoys.

Writing multiple-choice questions requires mastery of language and sentence structure. The responses must be appropriate to the stem and written in parallel structure. The wording should not provide grammatical cues or other clues to the correct answer.

Objective multiple-choice items can be structured as a direct question or incomplete statement. A direct-question format may be illustrated as follows:

Which of the following types of reliability is an underestimate of the true reliability?

a) Alternate forms
b) Kuder-Richardson
c) Split-half
d) Test-retest

The following is an example of incomplete statement format:

A college admissions director correlates scores from the SAT with students' third-year GPA. This is an example of

a) predictive validity
b) content validity
c) construct validity
d) concurrent validity

The direct question format is a more natural way to present information. However, a combination of formats provides variety and interest.

Assembling the Test

Test assembly includes packaging and reproduction. Kubiszyn and Borich offer some packaging tips:

Grouping together items of similar format,

FIGURE 3
Table of Specifications
Unit III Criteria of Satisfactory Measuring Instruments

Objectives	K	C	A	E	Total	Percent
Content						
A. Validity		2		1	3	
1. Content	2		2	1	5	
2. Concurrent	1	1	1		3	
3. Predictive	3	2	2	2	9	
4. Construct	1		1		2	
5. Coefficient of Correlation		1			1	
					23	51%
B. Reliability				1	1	
1. Test-Retest			1	1	2	
2. Alternative Forms		2	1		3	
3. Internal Consistency			1		1	
a. Split half		1	2		3	
b. Kuder-Richardson	1		1		2	
					12	27%
C. Accuracy						
1. Sources of Error		2			2	
2. Standard Error of Measurement		4	1		5	
3. Band Interpretation		1	2		3	
					10	22%
Total	8	16	15	6	45	100%
Percent	18	36	33	13		100%

Picture
Removed

arranging test items from easy to hard, properly spacing items, keeping items and options on the same page, placing illustrations near the descriptive material, checking for randomness in the answer key, deciding how students will record their answers, providing space for the test taker's name and the date, checking test directions for clarity, and proofreading the test before you reproduce and distribute it.[11]

Check your printer cartridge or typewriter ribbon to be sure that the master copy will reproduce well, and use good quality paper to make photocopies.

Administering the Test

Two types of preparation are required before you administer the test: physical and psychological. Physical preparation includes classroom seating arrangements, noise control, lighting, and temperature. A classroom that is crowded, noisy, too cold or too warm can affect students' test performance, thereby decreasing the accuracy of assessment.

Psychological preparation refers to the emotional climate provided by the teacher. Kubiszyn and Borich[12] offer some suggestions for test preparation:

1. *Maintain a positive attitude.* Classroom testing is intended to evaluate achievement and instructional procedures and to provide feedback for teachers and students. Hence, tests should not be used indiscriminately or to punish the class for uncompleted assignments. "Trick questions" should not be used.

2. *Maximize student motivation.* Encourage each student to do his or her best. Avoid comments that might impair students' test performance. Refrain from exaggeration or misleading statements about the test.

3. *Equalize advantages.* Help students become test-wise by reminding them of general test-taking strategies, such as these:

- Don't spend too much time on any difficult item.
- Try all items, then return to those you are unsure of.
- Check your answers for accuracy before turning in the test.

4. *Avoid surprises.* Students tend to perform better if they have sufficient time to prepare for a test, and know what material will be tested. Review and emphasize in class important concepts that will be included on the test.

5. *Clarify the rules.* Before distributing the tests, discuss time limits, and any special information about the answer sheets.

6. *Rotate distribution* so the same person is not always last to receive the test.

7. *Remind the students to check the pages and item numbers* to make sure nothing has been omitted.

8. *Monitor students.* Leaving the room during a test could be interpreted as an opportunity to compare answers or cheat.

9. *Minimize distractions.*

10. *Give time warnings.*

11. *Have a uniform policy for collecting tests.* This saves time and minimizes lost papers.

Scoring

Scoring a multiple choice test is relatively simple, compared to essays or short answer exams. Guidelines include the following:

1. Prepare a master answer sheet at the time you construct the test. Double check it to ensure accuracy.

2. Arrange the pages so that the test can be scored without the student being identified. (For example, use the first page for instructions and pupil identification. Fold back this sheet before scoring the test).

3. Consider using an answer sheet with an overlay or mechanical scoring aid to save time in grading.

Analyzing the Test

The quality of the test should be analyzed both before and after it is administered.

There are two kinds of analysis—qualitative and quantitative. *Qualitative analysis* means looking at test questions in terms of objectives, content validity, and technical item quality, appropriateness of response alternatives to the stem, and grammatical construction of the items. These should be checked before the test is administered. *Quantitative item analysis* is a numerical method for analyzing test items based on student responses. It uses item analysis to identify deficient questions, thus allowing the instructor to improve test quality. This analysis must be done after the test has been graded.

Once the tests have been scored, arrange the papers from highest to lowest score, and group them in two piles, the first containing the top half of the scores, the other containing the lower scores. For each item, calculate how many students from the two groups chose the various possible answers, and record the results on a form like the following (class size = 32):

Question 17

Options	Upper Half	Lower Half
A (correct answer)	9	6
B	4	3
C	3	1
D	0	3
E	0	3

From item analysis data compute two indices: difficulty index (D) and discrimination index (r). Use the following formula for the difficulty index (D):

D = *Number of students who answered the question right* ÷ *Total number of students in the class*

The item analysis shows that 15 students (9 from the upper half and 6 from the lower half) answered the item correctly out of 32 students in the class. Inserting these values in the formula above gives us a difficulty index of 47 percent.

$$D = \frac{15}{32} = 47 \text{ percent}$$

Since only 47 percent of the class got the answer right, this item is relatively close to the 50 percent difficulty index. Experts recommend that most items range between 20 percent and 80 percent in the difficulty index.

The discrimination index (r) from the item analysis data above can be computed using this formula:[13]

r = *Number of right responses in upper half – Number of right responses in lower half* ÷ *Number in each group*

$$r = \frac{9 - 6}{16} = .19$$

Some test construction experts insist on "r" being at least .30, while others claim that if it is positive, the item is adequate. In teacher-made tests, it is difficult to achieve discrimination indices above .30.

Returning the Test to Students

Unfortunately, final examinations are administered during the last class period, leaving no opportunity to go over the test with students. Concern for the quality of student evaluation should have caused you to review test items with your students after previous tests. This demonstrates your desire for feedback and willingness to make appropriate modifications in your tests.

Kubiszyn and Borich[14] suggest some debriefing guidelines to use *before* handing back tests and answer sheets:

1. *Discuss problem items.* Students will pay more attention to a discussion than if they are looking over their answer sheets.

2. *Listen to student reactions.* This reassures them that you are interested in their feedback and want to increase the validity and reliability of your test.

3. *Avoid on-the-spot decisions.* Tell the students that you will consider their comments, complaints, and suggestions, but that you will need time to study the test data.

4. *Be equitable.* Assure students that any scoring changes will apply to all students, not just those who raise objections.

After returning tests and answer sheets, do the following:

1. *Ask students to double check their tests.* It is not a sign of weakness to admit that you can make a mistake. Ask those who find errors to see you as soon as possible.

2. *Ask students to identify problems.* They will feel relieved at being given an opportunity to identify and discuss problem items. Hearing their point of view and clarifying muddy areas can be a helpful learning experience for both teacher and student.

Writing multiple-choice questions requires mastery of language and sentence structure.

Conclusion

The foregoing discussion makes it clear that reliable and valid tests do not just happen. Careful planning is necessary. Planning and constructing appropriate items can take up to 40 percent of class preparation time.

Testing is an integral part of instruction, although it should never be the only source of information about student achievement. Tests can enhance the educational process. Up to 25 percent of instructional time can profitably be used for testing. However, in the hands of ill-trained or inexperienced users, tests can be hazardous.

There is considerable room for improvement in measurement and evaluation in higher education. This can be accomplished through a two-pronged approach: (1) professional reading by each instructor, and (2) faculty development on the departmental level. Numerous books and periodicals can be helpful to teachers seeking to improve their testing and evaluation skills. Excellent results have been obtained from faculty development programs that feature in-service training in measurement evaluation. Such programs should include participants from various disciplines and the measurement community.

Now officially retired, Dr. Benjamin E. Bandiola was formerly Chairman of the Department of Education and Psychology at Southern College of Seventh-day Adventists, Collegedale, Tennessee. He currently serves as Adjunct Professor of Psychology for the University of Tennessee at Chattanooga.

NOTES AND REFERENCES

1. State teacher-education reform programs call for a course in tests and measurements.

2. NCATE, *NCATE Approved Curriculum Guidelines* (Washington, D.C.: NCATE, 1987).

3. Helpful publications relating to testing and measurement include the following: Anne Anastasi, *Psychological Testing*, 5th ed. (New York: Macmillan, 1982); Robert L. Thorndike and Elizabeth P. Hagen, *Measurement and Evaluation in Psychology and Education* (New York: Wiley, 1977); Robert L. Thorndike (ed.) *Educational Measurements* (Washington, D.C.: American Council on Education, 1971).

4. Edward J. Furst, *Constructing Evaluation Instruments* (New York: Longmans, Green and Co., 1958), p. 3.

5. Benjamin Bloom, M. Englehart, W. Furst, and D. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook 1: Cognitive Domain* (New York: Longmans Green and Co., 1956).

6. Ellen G. White, *Education* (Mountain View, Calif.: Pacific Press Publishing Association, 1903), p. 13.

7. David Krathwohl, Benjamin S. Bloom, and B. B. Masia, *Taxonomy of Educational Objectives, Handbook II: Affective Domain* (New York: McKay, 1964).

8. E. J. Simpson, *The Classification of Educational Objectives: Psychomotor Domain* (Urbana, Ill.: University of Illinois Press, 1971).

9. Tom Kubiszyn and Gary Borich, *Educational Testing and Measurement*, 3rd ed. (Glenview, Ill.: Scott, Foresman and Co., 1990), p. 54. Reprinted by permission.

10. White, *Education*, p. 232.

11. Kubiszyn and Borich, p. 116.

12. Ibid., pp. 119-121.

13. The symbol for coefficient of correlation "r" is used for discrimination index because it describes the relationship between the score on a test item and score on the total test for each examinee.

14. Kubiszyn and Borich, pp. 131, 132.