TEACHING TEST CONSTRUCTION

AND/OR VERSUS STATISTICS

E. Caldwell/University of South Florida

Introduction

Testing what students learn in school and college is largely a sham. That does not mean the practice is defective or unacceptable, because it has been the standard and accepted practice for generations. Testing tends to be a sham because, while we talk and write about the complex goals in learning, we measure only the most simple goals, such as recall of names, places, and definitions.

Testing has achieved a somewhat unsavory reputation because so many students experience tests in school that are characterized by "trickiness and triviality." There is considerable evidence to show that students direct their study toward what is expected on course examinations; i.e., they study for the test rather than to learn. Hence, the use of low-level measuring devices leads to the same level of learning.

Consider the following excerpt from a teacher-made test:

Test for Sixth Grade Social Studies

1. Lhasa is located on a _____ (stream or lake). (Page 368)
2. _____ was the capital of _____. (Page 369)
3. It was the _____ _____ of the religion known as _____. This is a branch of _____, a religion that began in _____. (Page 369)

The teacher who provided these items remarked, "This is the kind of test I had in school and it is like that used by most teachers I know." It is not clear what the test measures, but the level of learning represented is not very high.

The next illustration belongs to a special category that might be titled, "The Hour before Class Test" because it reflects little planning or skill in test development.

Science Test for Seventh-Eighth Grades

DIRECTIONS: Match the terms to the descriptions.

A. Conservation
B. Humus
C. Contour Planting
D. Fertilizer
E. Natural Resource
F. Depletion
G. —

1. Valuable materials found on and in the earth
2. Decayed plant and aminal products
3. Method of controlling depletion
4. Using up of minerals is called
5. Fuel
6. Minerals
7. —

True-False

1. Man belongs to the animal kingdom.
2. Man makes greater use of his abilities than all other animals.
3. The zebra is the "Ship of the Desert."

Completion

1. Where did the German Shepherd originate? _____
2. What dog is valuable in showing hunters where a bird is hiding? _____
3. What breed of dog is Lassie? _____

Test items like those above seem to be the rule rather than the exception. They ask students only to remember some of what has been read, what things are called, or how things are defined. The tests do not ask students to determine relationships, how and/or why things are related. The test items seldom call for the application of intellectual skills; they call only for recall.

The basic problem lies in the fact that few teachers receive any systematic or intensive training in test construction prior to entering the profession. If they have received some formal instruction, it was probably in a course titled something like, "Introduction to Educational Measurement." This is typically a course that includes some instruction in test construction, some statistical analysis, and the review of the technical qualities of a variety of standardized tests. Hence, the course doesn't provide much of a solution to the problem of poor classroom testing since so little time is devoted to the actual construction of tests.

The writer offers at least a partial solution to the

problem through the presentation of a course in which a major emphasis is placed on the development of test construction skills. While the enrollment is limited to students at the University of South Florida, the model would generalize to any university or in-service setting.

Method

For several terms, all students who have enrolled in the introductory measurement course taught by the writer have undertaken a major project. The purpose of the project is to help the students learn the various aspects of classroom test construction and to develop some of the skills required in test construction. The importance of the project is emphasized by the fact that successful completion of this project accounts for one-third of the student's final course grade.

Part one of the project consists of the construction of a classroom test for a particular group of pupils. The pupils may be those whom the student is teaching, since many of the students in the class are practicing teachers, or some other group of pupils. Arrangements are worked out so that the test to be constructed will be important to the pupils who take it.

Instruction is given to the students in the basic techniques associated with item writing. They are also introduced to *The Taxonomy of Educational Objectives: Cognitive Domain* (Bloom, 1956). Following this instruction, the students begin writing items designed to measure cognitive skills at each of the four levels: knowledge, comprehension, application, and analysis, as defined in the *Taxonomy*.

Part two of the project consists of the administration of the test to the selected group of pupils. This "field test" of the items is an important aspect of a test construction project, but it is seldom used in practice.

In the final part of the project, the student calculates all of the statistical indices that are taught in the course. Included are measures of central tendency and variability, item analysis statistics, converted scores, and inter-correlations of any subtests. All of the work is then narratively reported. Students use the narrative report to make judgments and claims about various qualities of their tests. Because a significant part of the project entails the application, calculation, and interpretation of statistical or other qualtitative functions, that part of the course receives considerable emphasis.

Results

Students have developed tests in many subjects and for many different age-ability levels. In most cases they have used the published taxonomies of intellectual skills, but in some cases modifications of, or new, taxonomies were developed. Tests have been constructed in music, various physical education areas, nursing, and mechanics, to name a few. Attempts to construct tests in the affective and psychomotor domains have not been fruitful. The most satisfying results came from tests designed to measure some complex intellectual skills in very young children.

The project appears to have advanced test construction skills of teachers in four specific ways, which are enumerated and illustrated:

1. The format of questioning is changed from random and aimless statements, lists of words, and terms to that of more direct and purposeful problem situations. For example:

Old Item:

(T-F) Man makes greater use of his abilities than all other animals.

New Item:

Does man depend on bacteria to help him?

A. No, bacteria causes man to get sick.
B. No, bacteria are used only in outer space.
C. Yes, bacteria are sometimes used to make aspirin and other medicines.
D. Yes, bacteria are used to decay materials not needed by man.

2. Both the teacher and pupil are encouraged to think beyond the simplest communication forms.

Old Item:

(T-F) Man makes greater use of his abilities than all other animals.

New Item:

Why is man considered smarter than all other animals?

A. Because he is larger.
B. Because he has learned to use plants and animals for purposes other than food.
C. Because he was made by God.
D. Because he stands up on his feet rather than crawling.

3. The computation of quantative descriptors such as item difficulty, item discrimination, and

reliability coefficients for one's own work increases the meaningfulness of the indices and the probability that the indices will be applied in real teaching/learning situations.

4. Tests that measure and stimulate abilities higher than memory-recall can be developed for very young children. The evidence for this is limited to face or logical validity, but indices such as difficulty and discrimination provide some support for this claim. Excerpts from two such tests that were constructed for third-grade level children illustrate this potential:

Example 1

Menu

Ice Cream . .	7¢	Soup	9¢
Coke	5¢	Hot Dog . .	10¢
Gum	4¢	Milk	8¢

Mary bought a coke and soup. Billy bought a hot dog and milk. Which sentence is true?

A. Mary's meal cost the most.
B. Billy's meal cost the most.
C. Billy and Mary spent the same amount.

How much change would you get back if you bought milk and gave the storekeeper a dime?

A. 17¢ B. 2¢ C. None

Example 2

Story

Susan wanted to make something. She was hungry for something sweet to eat. She got out her mother's cookbook from the drawer next to the sink. She turned to the section called "Desserts." She found a recipe that sounded good. Then she checked the ingredients to make sure they were all available. She followed the directions: First, she stirred one cup of butter until it was soft. Next, she blended two cups of sugar with the soft butter. Next, she added two eggs and two teaspoons of vanilla. When all of this was mixed well, she added three cups of flour, two teaspoons of baking powder, and one teaspoon of salt. Finally, she mixed in one cup of milk. Then it was time to pour the batter into some pans and bake it. She set the oven to 350° and put the pans full of batter in to bake. Thirty minutes later, Susan's dessert was done and ready for the icing.

Knowledge: What was the last thing Susan added?
A. Flour B. Milk C. Vanilla D. Baking Powder

Comprehension: The dessert Susan made was probably _____.
A. Pudding B. Ice Cream C. Cake D. Pie

Application: If Susan wanted to make twice as much dessert, how much baking powder would she need?
A. three teaspoons C. five teaspoons
B. four teaspoons D. six teaspoons

Analysis: What would happen if Susan left the dessert in the oven for an hour?
A. The dessert would be chewy.
B. The dessert would be delicious.
C. The dessert would fall apart.
D. The dessert would be too done.

In summary, classroom testing leaves much to be desired. Most measurement appears to occur at the lowest level of cognitive skills. Testing can be improved through the proper introduction to and instruction in test construction. The type of course described herein could help improve classroom testing.

Reference

Bloom, B. S. (Ed.), *Taxonomy of Educational Objectives,* Handbook 1, Cognitive Domain, New York: McKay, 1956.