Standards Are Needed

ONE OF THE newest testing instruments to evolve in American education is the criterion-referenced test (CRT),[1] believed by many educators to be a significant breakthrough in education. For many years, the watchword on the national education scene was "an equal education for all." Recently, American education has focused on maximizing the potential of each individual student. Developed in response to this new thrust, the CRT achieved almost immediate and widespread acceptance. This initial, broad acceptance by educators seems to assure a permanent place for the CRT in American education. However, such rapidly achieved status has resulted in an urgent need for establishment of standards both for the development of CRT's and for demonstrating their validity.

The latest edition of *Standards for Educational and Psychological Tests* of the American Psychological Association (APA) gives passing mention to CRT's in a short section on content-referenced and criterion-referenced *interpretation*, but the Standards apply principally to tests with norm-referenced interpretation. Many test publishers and educators feel strongly that APA Standards addressed solely to CRT development and validity are needed.

Many test publishers and educators feel strongly that APA Standards addressed solely to criterion-referenced test (CRT) development and validity are needed.

Criterion-referenced measurement is not a new or complicated concept. Teachers have traditionally had to plan a curriculum, teach to the broad goals of that curriculum, and develop and administer tests at appropriate stages in the instructional process to measure student progress. As any teacher will attest, the process of determining individual needs and providing instruction to meet those needs is an exacting one. Criterion-referenced instruction is simply a more clearly specified application of the traditional teaching method. A well-constructed CRT program, through use of instructionally relevant test instruments, identifies those students who have not yet mastered particular objectives, identifies those areas of nonmastery, and facilitates instructional decisions.

[1] Although other names for these tests have been used, such as objectives-referenced tests, domain-referenced tests, and objectives-mastery tests, "criterion-referenced tests" seems to be the most commonly used term.

for CRT's!

DAVID N. EVANS*

A CRT assesses an individual student's mastery or nonmastery of each objective in a set of explicit educational objectives. Every item written for the test is directly associated with an objective, which is usually stated in performance terms.

What Are the Constraints?

The selection of objectives for any given CRT is dependent on the constraints of test length and time, the curriculum needs of the educational community, and the nature of the population to be tested. Detailed analysis of a set of objectives is also required to identify and eliminate those objectives that may be implicit in others. A set of well-defined and comprehensive objectives can define a curriculum.

The development of a valid and comprehensive CRT is a long and demanding process. In a major CRT project, reading specialists spent a year and a half in the development of an initial list of more than 1,200 behavioral objectives, and this was only the beginning. Exhaustive study and analysis were required to specify such test characteristics as test levels, continuity between levels, number of items per objective, vocabulary, and item formats. The effort devoted to these and many other technical considerations culminated in the development of a two-stage item tryout that was administered to 18,000 students in grades 1 through 6 across the country.

The first stage of the tryout test program was a preinstruction administration of the items. This was then followed by a period of 10 to 12 weeks for instruction to be given in the normal curriculum of the participating schools. The second stage, a postinstruction test of the items, was administered to the same students. In the meantime, data were collected concerning whether an effort had been made to teach to each of the objectives included in the CRT.

The tryout program provided data on sensitivity to instruction, an important aspect of both the items and objectives. A procedure similar to that described by Marks and Noll [2] was used to reveal those items that were sensitive to instruction. As a measure of instructional validity, sensitivity to instruction data can validate individual items but does not explicitly invalidate items. However, where some items for a particular objective are demonstrated to be valid, those for the same objective that are not sensitive should probably be rejected.

Other purposes of the tryout test were: (a) to conduct standard item analyses, (b) to determine the minimum number of test items required for each objective, and (c) to determine the number of items a student must answer correctly to demonstrate mastery of an objective.

The preceding description necessarily touches only on the broad major considerations of CRT development. In the absence of definitive standards and specifications, test constructors must conscientiously provide, and teacher/administrators must demand, adequate supportive evidence of instructional validity of their CRT's. Sensi-

* *David N. Evans, General Manager, CTB/McGraw-Hill, Monterey, California*

[2] E. Marks and G. A. Noll. "Procedures and Criteria for Evaluating Reading and Listening Comprehension Tests." *Educational and Psychological Measurement* 27: 335-48; 1967.

. . . examinations were a great trial to me. The subjects which were dearest to the examiners, were almost invariably those I fancied least. . . . I should have liked to be asked what I knew. They always tried to ask what I did not know. When I would willingly have displayed my knowledge, they sought to expose my ignorance. This sort of treatment had only one result: I did not do well in examinations.—Winston S. Churchill. **My Early Life.** *New York: Scribner's, 1930.*

tivity to instruction is only one of several approaches to determine instructional validity. At present, however, it appears to be the most effective practical approach. In any case, definitive standards for the development of CRT's and for demonstrating their validity are needed.

A Call for CRT Standards

American educators have been expressing this need for CRT Standards for some time. In a paper presented at the joint APA-AERA-NCME Open Hearings on Standards held in San Francisco in March 1973, Glenn Roudabush, Senior Research Psychologist at CTB/McGraw-Hill, suggested certain revisions and additions to the APA Standards to accommodate criterion-referenced tests. One recommendation that he considered essential related to a new standard concerning the development of CRT's:

> When the test is a criterion-referenced achievement test, the accompanying manual should contain the full list of instructional objectives represented in (or measured by) the test, written out in full. It should also show the number of items used to measure each objective and, if a criterion or cut score is used in scoring the test in order to indicate mastery or nonmastery of the objectives, these cut scores should also be given.

Another recommendation acknowledged the importance of sensitivity to instruction in demonstrating construct validity:

> For criterion-referenced achievement tests, the construct can be thought of as the curriculum for which the test is intended to supply performance information. For this kind of test, then, appropriate construct validation would consist in studies demonstrating that performance on the test is sensitive to instruction related to the underlying curriculum, at least to the extent that that curriculum is fully represented in the set of instructional objectives upon which the test is based.

These proposed standards are not necessarily definitive; alternative procedures and standards for the development of CRT's may be suggested by concerned educators. For example, William E. Coffman of the University of Iowa, currently a member of the Analysis Advisory Committee of the National Assessment of Educational Progress, recently questioned whether sensitivity to instruction should be required for CRT items.[3] Also, item selection procedures for CRT's were debated in one of the leading journals on educational measurement.[4] An educational development with the enormous potential of the criterion-referenced test is bound to generate strong and diverse opinions.

It is time to air these opinions, to discuss theories of CRT development and validation, and to establish standards that will inspire the confidence of the entire American educational community. Only then will the CRT develop to its full potential as the most powerful tool of individualized instruction yet designed. □

[3] William E. Coffman. "A Moratorium? What Kind?" *Measurement in Education* 5 (2); Spring 1974.

[4] *Journal of Educational Measurement* 2 (2): 137-40; Summer 1974.