# Criterion-Referenced Measurement: Some Recent Developments

***CHARLES D. DZIUBAN***
***KENNETH V. VICKERY***

DURING the past decade, the notion of criterion-referenced measurement has been popularized in journals and at various scientific meetings. As a product of this movement, the criterion-referenced test (referred to as CRT) has been proposed as an alternative to the more traditional norm-referenced test (referred to as NRT).

In theory, a CRT is used to identify an individual's status with respect to a previously established standard of performance—possibly a behavioral objective. This is contrasted to an NRT in that an individual's performance must be identified with respect to the performance of others on the same test. This distinction has given rise to a substantial amount of research, speculation, and developmental work into the nature and consequences of criterion-referenced measurement.

Glaser (1963) proposed that a criterion-referenced measure was related to a student's acquisition of knowledge along a continuum ranging from no proficiency to perfect performance. He indicated that specific behaviors might be identified as standards for each level of knowledge. Later (Glaser and Nitko, 1971), that definition was clarified by stating that a criterion-referenced test is one constructed to yield measurements which are directly interpretable in terms of specified performance standards. Kriewall (1972) indicated that a CRT is one in which the items are homogeneous in difficulty for each examinee.

Emrick (1971) defined a CRT as not only having items of equivalent difficulty but equality in form and content as well. Carroll (1970) suggested that a CRT yields results which indicate as precisely as possible whether the pupil has achieved the specified goals of the learning task. Ivens (1970) defined a CRT as one composed of items keyed to a set of behavioral objectives. Livingston (1972) felt that it was sufficient to define a CRT as one for which a criterion score is specified without reference to the distribution of the scores in the group.

Continuing efforts have been made to delineate the characteristics of a CRT. Popham and Husek (1969) indicated that the basic difference between the two kinds of tests is in the concept of variability. Specifically, they indicated that variability is irrelevant to a CRT. Consequently, they speculated that teachers might encounter difficulty in the interpretation of traditional reliability and validity coefficients. It was

their suggestion that it is not possible, however, to tell a CRT from an NRT by looking at the two.

They stated that criterion-referenced tests are of two kinds, the ideal case and the more typical case. In the former situation the items are tied directly to the criterion so that the test is homogeneous in a very special way. That is, each individual who obtains the same score achieves it in essentially the same manner. This implies that if one knows an individual's score he also knows his response pattern. The possibility of such a device was previously discussed by Guttman (1944) and Tucker (1952). The second type of CRT is more typical, wherein the items are considered a sample from a potentially larger group. Knowing an individual's score may be largely indicative of his achievement, but it tells nothing about which items he missed.

Ivens (1970) attempted to develop and empirically evaluate a set of indices for criterion-referenced tests. These were intended to be used for the assessment of item and test quality. Since criterion-referenced tests may show a lack of variability, his indices were developed independently of test variance. He chose a set of ten behavioral objectives on general topics in probability, central tendency, and variability, and constructed six multiple-choice items for each of the ten objectives.

The entire item pool was administered as a pretest and post-test to students enrolled in a basic statistics course. From the item difference results (the difference between the number of correct responses for each item on the pretest and the post-test), he ranked the six items within each objective and constructed two tests. Form A consisted of two questions showing the greatest difference, that is, questions one and six for each objective. Form B was constructed from the items showing the least difference, items three and four for each objective. Parallel forms were constructed using the actual objectives as test items. The tests were administered as pretests, post-tests, and retests. Ivens concluded that item selection for criterion-referenced tests could be accomplished more efficiently using the difference in pretest–post-test item difficulty values.

Kriewall (1972) suggested that there were problems in applying norm-referenced metaphors to criterion test construction. Specifically, those were item difficulty, content validity, the assumption of normality, and test reliability. He asserted that teachers do not have a random sample of children but rather a particular group of individuals, and that inferences must be made concerning these children. He indicated that they must be treated as individuals, not as a random sample, and suggested four applications of criterion-referenced tests:

1. Categorize learners into temporary groups on the basis of a common requirement for instructional treatment (Diagnosis and Prescription Function).

2. Assess the relative effectiveness of competing instructional treatments (Instructional Assessment Function).

3. Determine, in the case of established instructional segments having predetermined performance standards, which individuals have acquired minimal standards of proficiency required for mastery and which learners require further prescriptive assistance (Quality Control Function).

4. In the case of curriculum development, indicate hierarchial relations within a content sequence (Curriculum Design Function).

Livingston (1972) proposed a reliability coefficient appropriate for criterion-referenced tests. His index was based on the relation of the size of the reliability coefficient to the range of talent in the individuals. Harris (1972b) pointed out, however, that the generally larger Livingston coefficient does not imply a smaller standard error of measurement and does not necessarily imply a better determination of whether or not a true score falls below a criterion value.

Hively *et al.* (1970) proposed a scheme for writing criterion-referenced test items. In their system an item form was comprised of a complete set of rules for generating a domain of items. Cox and Vargas (1966) contrasted norm-referenced item analysis with the criterion version. They ranked the items

on two indices which they developed, a difference index and pretest–post-test difference index, and obtained relatively low correlations between them. Their results suggested that some items which would be considered effective for a CRT would be discarded in the norm-referenced version.

Popham (1971) described the development of a prototypic criterion test item. Harris (1972a) recently reported the development of an index of efficiency for fixed length mastery tests. Hofman (1972) reported the development of an "e" index of item efficiency which may be of potential import to criterion-referenced measurement. Millman (1972) reviewed procedures for determining the number of items needed on criterion-referenced tests. He developed methods of relating test length, required accuracy, and proficiency standards.

Attempts have been made to explore the application of a CRT. Coulson and Cogswell (1965) emphasized the need for criterion-referenced tests with individualized instruction. Glaser and Cox (1968) also emphasized the similar need when it is important to differentiate between those who have mastered the objectives and those who have not. Millman (1970) advocated the use of a criterion-referenced marking system for the reporting of student progress.

Block (1972) suggested that the maintenance of different standards will maximize student learning depending upon the criterion by which the learning is operationalized. Proger *et al.* (1972) indicated that the flexibility of a CRT made it appropriate for individualized instruction and evaluation of handicapped children. Cox and Sterrett (1970) suggested the possibility of incorporating a criterion-referenced test within the norm-referenced version.

The recent research in criterion-referenced measurement has resulted in substantial progress. Apparently many believe, and possibly rightly so, that the concept offers a great deal of promise for individualized instruction. It has been strongly advocated in several areas and in some cases to the exclusion of the classical norm-referenced test.

At present, however, the state of the art leaves some questions for the classroom teacher. How are teachers to make the transition from the more traditional practices and what are the consequences? Can present instructional material be adapted to criterion-referenced measurement? Will a new system ultimately result in substantial additional demands upon teachers, many of whom are presently operating on overcrowded schedules?

Additional groundwork seems necessary to establish a framework for criterion-referenced measurement in the classroom on a day-to-day basis. It seems appropriate that consideration be addressed to these topics even as the dialogue over the appropriate measurement model and efficiency indices persists.

It does seem true that norm-referenced metaphors may not be appropriate to criterion-referenced assessment. If, in fact, it is impossible to distinguish the two kinds of tests by inspection, it must be the results and how they are used which make the difference. The search, then, for analogs to traditional

indices of test effectiveness may not be fruitful. If a norm-referenced test results in a complete ordering of individuals, a criterion-referenced version must result in a partial ordering. It seems clear at present that much more developmental work is necessary if a CRT is to develop into a practical reality for the classroom teacher.

### References


James Block. "Student Evaluation: Toward the Setting of Rational, Criterion-Referenced Performance Standards." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972.

John Carroll. "Problems of Measurement Related to the Concept of Learning for Mastery." *Educational Horizons* 48: 71-80; 1970.

J. Coulson and J. F. Cogswell. "Effects of Individualized Instruction on Testing." *Journal of Educational Measurement* 2: 59-64; 1965.

Richard Cox and Barbara Sterrett. "A Model for Increasing the Meaning of Standardized Test Scores." *Journal of Educational Measurement* 7: 227-28; 1970.

Richard Cox and Julie Vargas. "A Comparison of Item Selection Techniques for Norm Referenced and Criterion Referenced Tests." Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, 1966.

John Emrick. "An Evaluation Model for Mastery Testing." *Journal of Educational Measurement* 8: 321-26; 1971.

Robert Glaser. "Instructional Technology and the Measurement of Learning Outcomes." *American Psychologist* 18: 519-21; 1963.

Robert Glaser and Richard Cox. "Criterion-Referenced Testing for the Measurement of Educational Outcomes." In: R. A. Weisgerber, editor. *Instructional Process and Media Innovation.* Chicago, Illinois: Rand McNally & Company, 1968.

Robert Glaser and Anthony Nitko. "Measurement in Learning and Instruction." In: R. L. Thorndike, editor. *Educational Measurement.* Washington, D.C.: American Council on Education, 1971.

Louis Guttman. "A Basis for Scaling Qualitative Ideas." *American Sociological Review* 9: 139-50; 1944.

Chester W. Harris. "An Index of Efficiency for Fixed-Length Mastery Tests." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972(a).

Chester W. Harris. "An Interpretation of Livingston's Reliability Coefficient for Criterion Referenced Tests." *Journal of Educational Measurement* 9: 27-29; 1972(b).

Wells Hively, II, *et al.* "Introduction to Domain-Referenced Achievement Testing." Symposium presentation, American Educational Research Association, Minneapolis, Minnesota, March 2-6, 1970.

Richard Hofman. "The Efficiency Index in Item Analysis." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972.

Stephen Ivens. "An Investigation of Item Analysis, Reliability, and Validity in Relation to Criterion Referenced Tests." Unpublished doctoral dissertation, Florida State University, 1970.

Thomas Kriewall. "Aspects and Applications of Criterion-Referenced Tests." Technical Paper #103. Presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972.

Samuel Livingston. "Criterion-Referenced Applications of Classical Test Theory." *Journal of Educational Measurement* 9: 13-26; 1972.

Jason Millman. "Passing Scores and Test Lengths for Domain Referenced Tests." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972.

Jason Millman. "Reporting Student Progress: A Case for a Criterion Referenced Marking System." *Phi Delta Kappan* 52: 226-30; 1970.

W. James Popham. "Indices of Adequacy for Criterion-Referenced Tests." In: W. James Popham, editor. *Criterion-Referenced Measurement, An Introduction.* Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

W. James Popham and T. R. Husek. "Implications of Criterion Referenced Measurement." *Journal of Educational Measurement* 6: 1-9; 1969.

Barton Proger, Lester Mann, Robert Burger, and Lawrence Cross. "Adapting Criterion-Referenced Measurement to Individualization of Instruction for Handicapped Children: Some Issues and a First Attempt." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972.

Leonard Tucker. "Scales Minimizing the Importance of Reference Groups." In: *Proceedings, Invitational Conference on Testing Problems.* Princeton, New Jersey: Educational Testing Service, 1952. pp. 22-28.



—CHARLES D. DZIUBAN, *Assistant Professor of Education, Florida Technological University, Orlando; and* KENNETH V. VICKERY, *Physics Teacher, Edgewater High School, Orlando, Florida.*