Contributions to Criterion-Referenced Testing Technology: An Introduction

Ronald K. Hambleton
University of Massachusetts, Amherst

Glaser (1963) and Popham and Husek (1969) were the first researchers to draw attention to the need for criterion-referenced tests, which were to be tests specifically designed to provide score information in relation to sets of well-defined objectives or competencies. They felt that test score information referenced to clearly specified domains of content was needed by (1) teachers for successfully monitoring student progress and diagnosing student instructional needs in objectives-based programs and by (2) evaluators for determining program effectiveness. Norm-referenced tests were not deemed appropriate for providing the necessary test score information.

Many definitions of criterion-referenced tests have been offered in the last 10 years (Gray, 1978; Nitko, 1980). In fact, Gray (1978) reported the existence of 57 different definitions. Popham's definition, reported by Hambleton (1981) in a slightly modified form, is probably the most widely used:

> A criterion-referenced test is constructed to assess the performance levels of examinees in relation to a set of well-defined objectives (or competencies).

Five points about the definition require explanation. First, terms such as objectives, competencies, and skills may be used interchangeably. Second, each of the objectives measured in a criterion-referenced test should be well defined. This means that the content or behaviors defining each objective must be clearly described. Well-defined objectives make the task of writing test items easier and improve the quality of test score interpretations. Item writing is easier because appropriate content is spelled out. The quality of interpretations is improved because of the clarity of the content or behavior domains to which test scores are referenced. Third, when more than one objective is measured in a test, usually the test items are organized into nonoverlapping subtests corresponding to the objectives, and examinee performance is reported on each of the objectives.

Fourth, the definition does *not* include a reference to a cutoff score or standard. It is common to set a standard of performance for each objective measured in a test and to interpret examinee performance in relation to it. However, descriptive interpretations of scores such as "Student A is estimated to have mastered 70% of the content measuring Objective 1" are also made, and standards

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 421–424

are *not* used in this type of score interpretation. That a standard need not be set on a criterion-referenced test may come as a surprise to some who have mistakenly assumed that the word *criterion* in *criterion-referenced test* refers to a standard or cutoff score. In fact, *criterion* is a word used by both Glaser (1963) and Popham and Husek (1969) to refer to a *domain of content or behavior* to which test scores can be referenced.

Fifth, psychometric and statistical models have been developed to address two different conceptualizations of mastery underlying criterion-referenced test score performance. A continuum model is based on the assumption that in relation to each objective, examinees may vary in terms of their true level of performance. Examinee true scores can vary across the total range of possible test scores. The problem, then, becomes one of choosing and applying a method for identifying a point on the score scale (referred to as a cutoff, or advancement, score) that can be used to separate examinees into two mastery states: *mastery* and *non-mastery*. A state model, on the other hand, is based on the assumption that in relation to each objective, examinees have either total mastery (100%) or no mastery (0%) of the content. Under this conceptualization of mastery, with a reasonable number of test items, the matter of setting a cutoff score on each objective is relatively easy because of the bimodal nature of the distribution of examinee scores.

Since the initial work by Glaser (1963) and Popham and Husek (1969), over 600 papers have been written on the topic of criterion-referenced testing (Hambleton, Swaminathan, Algina, & Coulson, 1978) and applications have been extended from the classroom to state-wide assessments, school promotion examinations, and professional licensure and certification examinations. Criterion-referenced tests are now widely used at all levels of public and private education and in industry and the military. The impact of criterion-referenced testing (or, as it is alternately referred to, domain-referenced testing, objectives-referenced testing, basic skills testing, competency testing, or performance-based testing) has been substantial, widespread, and significant.

Popham and Husek (1969) also offered a set of methods and procedures for constructing criterion-referenced tests and interpreting test scores. Following their pioneering work, many papers have been written on technical matters associated with building criterion-referenced tests. In fact, the psychometric literature abounds with papers on such topics as definitions, writing and selecting objectives, preparing and validating test items, determining test lengths, assembling tests, assessing reliability and validity of test scores and decisions, estimating examinee performance and making decisions, reporting scores, and evaluating tests.

Purpose of the Special Issue

In view of the considerable interest among psychometricians in the criterion-referenced testing field and in view of the large number of technical contributions that have appeared in recent years, many of them in *Applied Psychological Measurement (APM)*, the editor and editorial board of *APM* felt that a full issue devoted to recent technical developments would be of considerable value. It was hoped that this issue would inform readers of recent technical advances, would offer new conceptualizations of problems and results to further the growth of the field, and would suggest many promising directions for additional research.

The six papers in the special *APM* issue can be organized around three broad topics: Test Development (Wilcox); Test Score Uses (Shepard, van der Linden, Macready and Dayton); and Evaluation of Scores and Decisions (Traub and Rowley, Linn). Each of the papers addresses a technical area that has a central role in the development and proper use of criterion-referenced tests. The authors of the papers were encouraged to accomplish three goals:

1. To provide a review of relevant literature,
2. To offer new models and/or results (or new ways for viewing technical issues and problems), and
3. To suggest several promising directions for additional research.

Introduction to the Papers

In the first paper, Wilcox considers the problem of determining the number of items necessary to produce a criterion-referenced test with "desirable properties." Of course, the desirable properties must be chosen by the test developer, and these properties may vary from one testing situation to the next and from one developer to another. An example of a desirable property would be a lower bound figure on the probability of examinees being correctly classified. Wilcox considers several possible true scores that might be adopted by test developers. In addition, several promising solutions to the determination of test length problem, based upon sets of assumptions and desirable outcomes, are offered.

Standard-setting issues and methods are considered by Shepard in the second paper. In addition to considering the controversy surrounding the use of standards, she provides a comprehensive review of many of the more promising standard-setting methods and considers the selection of a standard-setting method in relation to three uses of criterion-referenced tests: diagnosis, certification, and program evaluation.

Van der Linden addresses the decision-theoretic problem of assigning examinees to mastery states in the third paper. He advances a number of statistical models and, like Wilcox, considers three true score estimates. Van der Linden addresses the problem of optimally selecting a cutoff score on a test score scale with respect to one of three loss functions (threshold, linear, and normal-ogive), with and without an external criterion measure.

In the fourth paper, Macready and Dayton advance a set of statistical models that can be used in situations where competency acquisition is viewed as "all-or-none." The Macready-Dayton state models and applications are not as well known as the continuum models reviewed in the van der Linden paper, but they do appear to have considerable potential when applied to appropriate types of content.

Perhaps no topic in the criterion-referenced testing field has attracted as much interest as the one considered by Traub and Rowley in the fifth paper. Traub and Rowley consider methods for assessing the reliability of test scores and decisions. They organize the available methods according to a consideration of two dimensions: type of test scores (dichotomous or continuous) and the intended use of the scores (to make descriptions or decisions). Most of their paper is focused on reliability methods associated with continuous scores being used to make decisions (typically binary) because nearly all of the reliability literature applies to this particular test situation.

Linn is the author of the sixth and final paper on issues of validity with criterion-referenced tests. Linn advances the position that content validation evidence is not sufficient evidence to insure the validity of intended score interpretations and decisions. He argues that descriptions and decisions are made on the basis of examinee test performance and, therefore, that validation evidence bearing on these intended uses of scores is necessary. Insights into what is required in suitable validation studies are gained from two examples.

It seemed desirable to conclude the special issue with some commentary on the six papers. Thus, Berk and Livingston have reviewed the papers and have provided some of their own views on future directions for technical developments.

Background References

Of the many journal articles, chapters, and books in the criterion-referenced testing literature, several publications will be of special interest to readers of this issue because they serve to define, to review, and/or to critique large segments of the expansive literature. The following publications will provide both background material and follow-up readings on topics not included in this issue.

Popham's (1978) book entitled, *Criterion-Referenced Measurement* provides an especially good entry point into the literature because Popham has reviewed the history of criterion-referenced testing, has compared norm-referenced and criterion-referenced tests, has provided directions for preparing well-defined objectives, and has introduced many of the technical issues and methods associated with criterion-referenced tests. Nitko (1980) has organized the many varieties of criterion-referenced tests (including one which was introduced in 1864), and has offered a very helpful classification scheme for tests. Nitko's classification scheme is based on a consideration of (1) the clarity of domains of content measured by tests and (2) whether or not content within domains can be ordered in logical ways.

Millman (1974) has reviewed a number of methods for preparing well-defined objectives, and has offered a set of steps for building criterion-referenced tests and evaluating test scores and decisions. The Millman paper has had a substantial impact on the field because of its clarity, quality of ideas, comprehensiveness, and suggestions for future research and developmental activities. Hambleton et al. (1978) provided a review of many of the technical contributions to the criterion-referenced testing field. Special attention has been given in their review to (1) statistical estimation of domain scores; (2) the uses of decision theory; and (3) the assessment of reliability.

Finally, the most up-to-date review of the criterion-referenced testing field has been provided in the collection of papers edited by Berk (1980). The book contains six papers presented at the first annual Johns Hopkins University National Symposium on Educational Research in Washington, DC, in 1978. The topics addressed by the authors are preparation of objectives (referred to as domain specifications), generation of test items, item analysis, test score validity, standard setting, and reliability assessment. In addition to providing comprehensive literature reviews, the authors have provided guidelines for improving the practice of criterion-referenced testing and have offered suggestions for additional research.

References

Berk, R. A. (Ed.) *Criterion-referenced measurement: State of the art.* Baltimore, MD: The Johns Hopkins University Press, 1980.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, *18*, 519-521.

Gray, W. M. A comparison of Piagetian theory and criterion-referenced measurement. *Review of Educational Research*, 1978, *48*, 223-249.

Hambleton, R. K. Advances in criterion-referenced testing technology. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology*. New York: Wiley, 1981.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, CA: McCutchan, 1974.

Nitko, A. J. Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research*, 1980, *50*, 461-485.

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Special Issue Editor's Address

Ronald K. Hambleton, University of Massachusetts, Laboratory of Psychometric and Evaluative Research, Hills South-Room 154, Amherst, MA 01002.