DOCUMENT RESUME

ED 360 367 TM 020 271

AUTHOR Hambleton, Ronald K.
TITLE The Rise and Fall of Criterion-Referenced Measurement?
PUB DATE Apr 93
NOTE 16p.; Paper presented at the Annual Meetings of the American Educational Research Association (Atlanta, GA, April 12-16, 1993) and the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Content Validity; *Criterion Referenced Tests; Decision Making; Diagnostic Tests; Educational Assessment; Educational History; Educational Trends; Elementary Secondary Education; *Evaluation Methods; Literature Reviews; *Measurement Techniques; *Research Methodology; Scholarly Journals; Student Evaluation; *Teacher Education; *Test Construction; Test Items
IDENTIFIERS *Performance Based Evaluation; Standard Setting; Test Specifications

ABSTRACT

A review of educational measurement journals indicates that the decade of the 1970s and up until about 1984 was the time when substantial numbers of studies dealt with criterion-referenced measurement, and the era in which criterion-referenced measurement advances were made. Although there appears to be less interest in publications today, a number of measurement advances can be traced directly to the conceptual paper of R. Glaser (1963) on the topic of criterion-referenced measurement. The following six areas are highlighted: (1) clarification in specifying performance outcomes; (2) improvements in item writing and increased emphasis on content validity; (3) new approaches to reliability and validity methods and proficiency estimation; (4) new and improved standard-setting methods; (5) increased emphasis on diagnosis, decision-making, and criterion-referenced interpretations; and (6) improved training of teachers in the area of assessment. Although less research is apparent in measurement journals, it is evident that the concept lives in the wealth of tests and measurement textbooks used today. Two figures illustrate the discussion (Contains 15 references.) (SLD)

The Rise and Fall of Criterion-Referenced Measurement?[1,2]

Ronald K. Hambleton
University of Massachusetts at Amherst

A review of the educational measurement journals would reveal that the decade of the 1970s and the period from 1980 to 1984 was the time when substantial numbers of criterion-referenced measurement advances were made. The journals were filled with contributions which expanded, clarified, and operationalized the concept of criterion-referenced measurement introduced by Professor Robert Glaser in 1963. After 1984, the papers, reports, and books stopped coming, or, more correctly, the numbers declined to a trickle. In their place came papers about latent traits, and Rasch and Birnbaum models. Item bias and test score equating studies became the rage, and more recently, authentic measurement, performance assessment, assessment of higher-order thinking skills, and portfolios have become the focus of attention for policy-makers, educators, and researchers. These latter topics appear to be taking educational measurement in a new direction. What's going on? Did the criterion-referenced measurement concept die with the publication of Ron Berk's edited book, A Guide to Criterion-Referenced Test Construction published in 1984?

Figure 1 highlights the trend in criterion-referenced measurement publications since 1963. The numbers reported in Figure 1 came from the

---

[1]Laboratory of Psychometric and Evaluative Research Report 256. Amherst, MA: University of Massachusetts, School of Education.

[2]Paper presented at a symposium entitled, "Criterion-Referenced Measurement: A Thirty Year Retrospection" at the joint meetings of NCME and AERA, Atlanta, April 14, 1993.

bibliography I use in my criterion-referenced measurement course at the University of Massachusetts, hence the numbers do not necessarily reflect the precise figures that might be found from a scientific survey of books, reports, and journals over the last 30 years. Not surprisingly, the bibliography overrepresents the contributions of the instructor of the course! Still, the figures are revealing. Over 70% of the articles in the bibliography appeared between 1975 and 1984. The rise and then the fall of criterion-referenced measurement papers in the testing literature is quite evident. Was it the perception of many educators in the middle 1980s that the utility of criterion-referenced tests was limited to the assessment of lower level or basic skills with multiple-choice test items and that their role, therefore, in the new educational reform movement, which began around 1985, would be limited?

The impact of Glaser's criterion-referenced measurement paper on research and educational testing practices for many years after publication was low. Textbooks on educational testing up until at least 1974 rarely gave more than a few pages to the topic of criterion-referenced measurement. Popular and important measurement journals such as the Journal of Educational Measurement published only a few papers: only thirteen articles between 1964 and 1974 - a period of 11 years. That's only about one per year. A well-known test developer at a very big testing company in this country told me that he had been assigned the task of watching criterion referenced testing until it died. The evidence available in 1975 for criterion-referenced testing was not encouraging. But a series of forces came together that led to a surge of interest in research, development, and applications of criterion-referenced measurement concepts that far surpassed what we have seen in recent years with item response theory. That may seem hard to believe given the

avalanche of IRT papers in our measurement journals but the reason is clear. Whereas item response models have been embraced by the educational measurement community, primarily, criterion-referenced measurement became popular with teachers and administrators, and specialists in curriculum and instruction, special education, and program evaluation, in addition to the measurement specialists working in district, state, and national testing agencies, not to mention the military, industry, and over 900 credentialing organizations. Beginning around 1975 the amount of research jumped up tremendously and there was evidence of criterion-referenced testing programs in nearly every school district and state department of education in the country. At one time, in the 1980s, 49 of the 50 states had criterion-referenced testing programs in place. Basic skills testing or minimum competency testing and program accountability were the forces that drove the interest in criterion-referenced measurement.

In the remainder of this paper, attention will be given to measurement advances that can be directly traced to Professor Glaser's conceptual paper on the topic of criterion-referenced measurement. The rise and fall of the publications are evident, but the impact of criterion-referenced measurement on testing and measurement practice to-day is broad, substantial, and quite easy to document. For the purposes of this paper, six areas will be highlighted. The areas are listed in Figure 2.

Though my task is to look at the impact of Glaser's paper on the field of testing 30 years later, a few words about Jim Popham and his seminal role is very much in order. Though Professor Popham may have taken his first marching orders from Professor Glaser, he quickly became one of the important national commanders. Jim Popham deserves a star beside his name for his continuous and important criterion-referenced measurement contributions for

nearly 25 years. His paper with the late Ted Husek (Popham & Husek, 1969) provided the measurement push to go along with Glaser's conceptual push that put the criterion-referenced testing ball in motion. The papers by Glaser (1963) and Popham and Husek (1970) were the two papers which provided the cornerstones for most of the criterion-referenced measurement work that followed.

Measurement Advances

Clarification in Specifying Performance Outcomes

One of the most important contributions of criterion-referenced measurement to testing practice was the central focus it placed on describing the intended outcomes of instruction, i.e. the objectives. Requiring teachers and test developers to clearly describe the knowledge and skills to be tested provided the framework needed to write valid test items, to evaluate item-objective congruence, and to enhance the quality of test score interpretations. Behavioral objectives, and later, amplified objectives and test specifications (see Popham, 1978), provided the framework needed to construct valid criterion-referenced testing systems. Glaser and Popham were correct when they argued for the need for clear statements of the content domains of interest. Today, item and test specifications are routinely prepared by persons implementing criterion-referenced testing programs. The impact, too, of test specification technology on performance assessment is clearly evident. Performance assessments are being developed to-day from very detailed specifications about domain definitions, testing procedures, and scoring.

Even item generating rules which have limited applicability and represent an extreme form of item specifications are finding their way into computer-based testing systems. When they are applicable, they can greatly facilitate the storage and generation of criterion-referenced tests. Jay

Millman's work and the earlier work by Wells Hively and his colleagues (Hively, Patterson, & Page, 1968) have been particularly influential on testing practices.

Improvements in Item-Writing and Increased Emphasis on Content Validity

With the advent of criterion-referenced testing, there was a central concern among test developers for item validity. Since test scores would be interpreted in terms of objectives, it was incumbent upon test developers to write test items which were technically sound but also which closely matched the objectives they were written to measure and were representative of the domains of content spanned by the objectives. The result was a greater focus on item writing training and item review to enhance the content validity of criterion-referenced tests. At one stage, content validity concerns were even seen as sufficient to justify the use of criterion-referenced tests though this position was changed as the field developed (Messick, 1975).

Certainly, too, the desire to match items to objectives resulted in more careful attention to item writing training and item reviews in the credentialling exam area. And though problems still remain in this area, there is substantial evidence to suggest that the criterion-referenced measurement movement left its impact on item writing and a commitment to content validation studies. Even to-day there is a strong sense that content validity evidence is sufficient to support the use of many credentialing exams. Methods and procedures for assessing content validity which were produced in the 1970s are in wide use to-day.

New Approaches to Reliability and Validity Methods and Proficiency Estimation

Whether or not criterion-referenced measurement had come along, test scores would certainly be used to sort examinees into ability groupings, or mastery categories. With the advent of criterion-referenced testing programs

there was a rush among psychometricians to reconceptualize the meaning of reliability and validity (see, for example, Hambleton & Novick, 1973; Hambleton, Swaminathan, Algina, & Coulson, 1978) and to enhance the accuracy of proficiency estimates. This work was necessary because many standard test theoretic concepts, models, and practices which had been developed to aid norm-referenced testing practice were not applicable to criterion-referenced testing practice. In addition, decisions from criterion-referenced test scores were often being made with limited amounts of information. This resulted in the application of decision theoretic concepts, beta-binomial models, generalizability theory, Bayesian statistical procedures and other methodologies to prepare new approaches to the assessment of reliability and validity and proficiency estimation. Some of this work, which was fostered and developed as part of the criterion-referenced measurement movement, remains to-day and statistical concepts such as decision consistency, kappa, decision accuracy, false-positive and false-negative error rates, remain integral parts of the technical literature on educational testing.

New and Improved Standard-Setting Methods

Despite the controversies that still surround standard-setting methods, these methods have advanced considerably since Glaser's 1963 paper. Spurred on by the desire of many users of criterion-referenced tests to distinguish between masters and non-masters, certifiable and non-certifiable, etc., advances in standard-setting have been directly due to the increase in interest and use of criterion-referenced tests. Professor Glaser never intended for criterion-referenced measurement to be so closely associated with standard setting and the use of mastery-non-mastery decisions but that has become the case. In fact, one of the great myths about criterion-referenced measurement was that the word "criterion" meant standards. Glaser himself

seemed much more interested in developing psychologically interpretable scales along which persons could be positioned.

Before 1963, the Nedelsky standard-setting method, which applied to multiple-choice items only, was the one method in the measurement literature (Nedelsky, 1954). Also, normative based methods, which directly established the passing rate, existed and were popular in the credentialing field. The Angoff method was introduced in 1971 as a footnote, perhaps even an after-thought, in the main study of the day on equating (Angoff, 1971). But, because of the popularity of criterion-referenced tests, the 1970s and 1980s saw the development of many new standard setting methods, research on these methods, guidelines for applying the methods, and specific methodological studies aimed at issues such as the number of judges, their training, the role of performance data, etc. (see, for example, Berk, 1986; Jaeger, 1989; Shepard, 1984). All of this work has served the measurement field well and is being used in school districts, state departments of education, and national credentialing agencies. Certainly the measurement field has come a long way since the days when the passing scores were literally pulled "from the air" or set to insure a particular failure rate. Even the new standard-setting methods which are being applied today to performance assessments and portfolios (e.g., the modified Angoff method) have their roots in the criterion-referenced testing movement.

Increased Emphasis on Diagnosis, Decision-Making, and Criterion-Referenced Interpretations

One of the most positive aspects of the criterion-referenced measurement movement is that it has focused attention on diagnoses of learning deficiencies. Criterion-referenced tests are being used successfully in many places and with many types of students to identify strengths and weaknesses.

Criterion-referenced tests, too, have focused more attention on the use of tests to make instructional decisions. Persons are looked at in relation to well-defined areas of content and decisions or actions are taken. Finally, probably every first year teacher to-day knows the fundamental distinction between norm-referenced and criterion-referenced interpretations of test scores. Focusing attention on persons and the skills and knowledge they are responsible for or need to know has provided teachers with an attractive option to comparing students with one another and grading on the curve. The focus of assessment on the examinee and the body of knowledge and skills he/she is responsible for is criterion-referenced and is central to the role of assessment in today's reform movement characterized by performance assessments and portfolios.

Improved Training of Teachers in the Area of Assessment

A review of the new AFT, NCME, and NEA teacher competencies in the educational assessment of students (AFT, NCME, NEA, 1990) show clearly the impact of criterion-referenced testing. Nearly every standard in one way or another addresses the design, administration, evaluation, and interpretation of criterion-referenced tests. Clearly, the two teacher organizations seem very comfortable with the concepts and methods of criterion-referenced measurement and want to insure that teachers are well-trained in this area. The impact of criterion-referenced measurement is quite clear and positive, especially so when the views of the AFT and NEA about criterion-referenced testing are compared to their views about norm-referenced testing.

Conclusions

Have we witnessed the rise and fall of criterion-referenced measurement? Absolutely not! The rise is clearly documented in Figure 1. The period from

1970 to 1984 was a great period for advances in criterion-referenced measurement and important model-building, technical developments, and application work. And since 1984? It's true that there is less research in the journals but there is ample evidence that the concept lives. It lives in the wealth of tests and measurement textbooks that are to-day being used to train the next generation of teachers and test developers. In fact many textbooks give nearly equal treatment to norm-referenced and criterion-referenced testing (see, for example, Crocker & Algina, 1986). It lives in the 900+ credentialling organizations who now use various parts of the technology associated with criterion-referenced tests to build and report test scores and decisions. It lives in the numerous state departments of education that continue to construct, administer, and use criterion-referenced tests in individual and program assessment. Most certainly the concept lives in the authentic measurement and performance assessment movements. These are very much movements centered in the criterion-referenced measurement framework. Individuals are assessed in relation to well-defined outcomes of instruction and well-defined standards or expectations. Certainly the concept of criterion-referenced measurement is not limited to low level basic skills and multiple-choice test items. It lives in large chunks of military job performance testing such as the Army's skill qualification tests. I could go on.

So, in sum, the field of cognitive measurement has probably been redirected for ever. We now have two important frameworks, norm-referenced and criterion-referenced frameworks, for test development and test use. It is hard to find fault with criterion-referenced assessment when it is focused on what persons are expected to know and to be able to do, when assessment is person-centered as opposed to group centered, and when substantial commitments

are made to insuring that the assessments themselves are valid, and care is given to the interpretation of results. Thanks to Professor Glaser's insights, wisdom, and contributions, we have had a 30-year start on making educational measurements more meaningful to persons who must make important decisions about persons and programs.

References

American Federation of Teachers, National Council on Measurement in Education, National Educational Association. (1990). Standards for teacher competence in educational assessment of students. Educational Measurement: Issues and Practice, 9(4), 30-32.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.; pp. 508-600). Washington, DC: American Council on Education.

Berk, R. K. (Ed.). (1984). A guide to criterion-referenced test construction. Baltimore, MD: The Johns Hopkins Press.

Berk, R. K. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56(1), 137-172.

Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. New York: Holt, Rinehart, and Winston.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-521.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10(3), 159-170.

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

Hively, W., Patterson, H.L., & Page, S. A. (1968). A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 5, 275-290.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed.; pp. 485-514). New York: Macmillan.

Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

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

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

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169-198). Baltimore, MD: The Johns Hopkins Press.

Figure1. Criterion-Referenced Measurement Contributions Organized By Publication Date

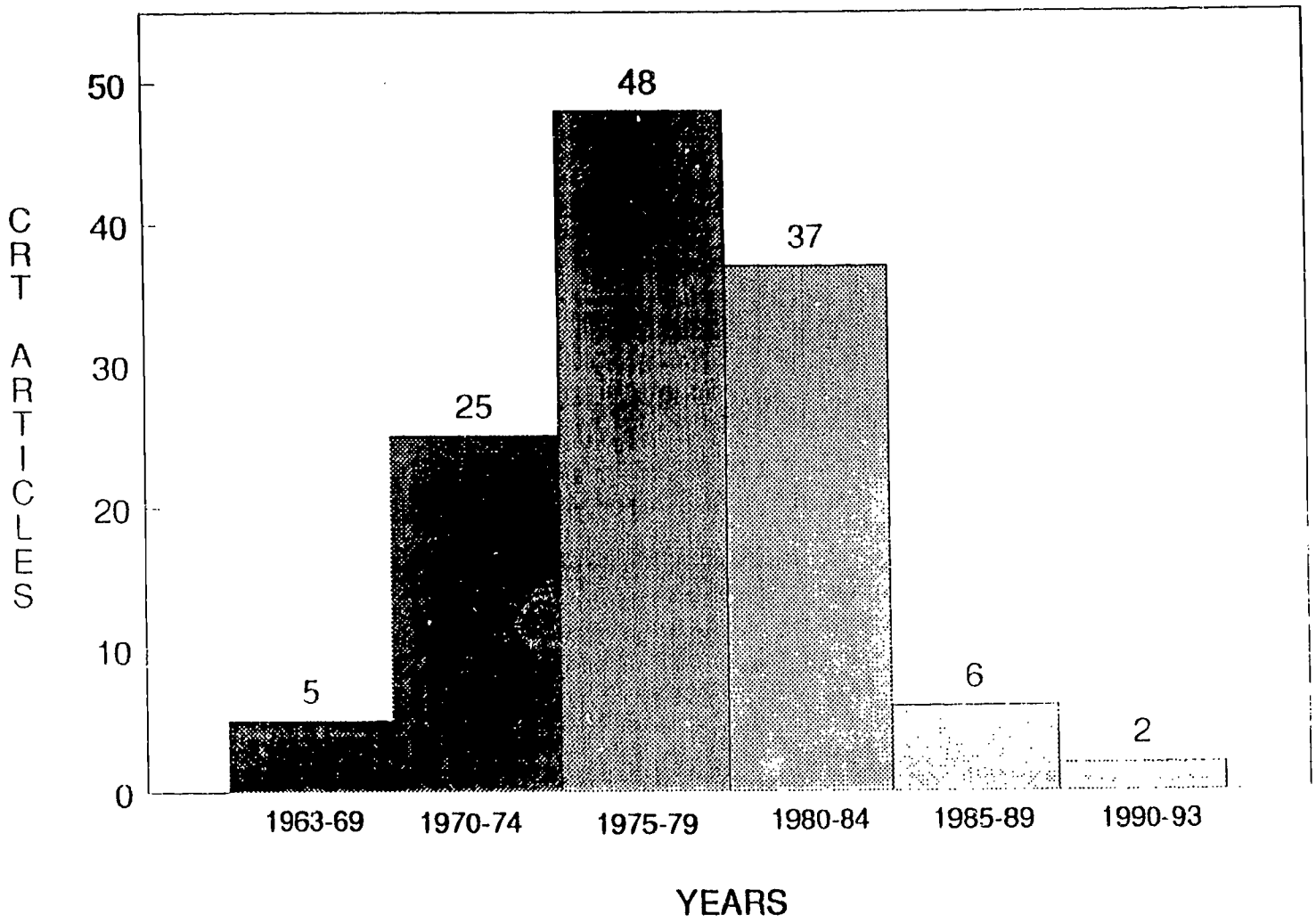

Figure 2. Criterion-referenced measurement advances which have impacted on educational testing practices.

a. Clarification in Specifying Performance Outcomes

b. Improvements in Item Writing and Increased Emphasis on Content Validity

c. New Approaches to Reliability and Validity Methods and Proficiency Estimation

d. New and Improved Standard Setting Methods

e. Increased Emphasis on Diagnoses, Decision-Making, Criterion-Referenced Interpretations

f. Improved Training of Teachers in the Area of Assessment