DOCUMENT RESUME

ED 189 183 TM 800 365

AUTHOR Nitko, Anthony J.
TITLE Individual Differences Between Criterion-Referenced Tests. Working Paper #39 of the Program in Educational Research.
INSTITUTION Pittsburgh Univ., Pa. School of Education.
PUB DATE Apr 80
NOTE 18p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).
AVAILABLE FROM Educational Research Department, 5C03 Forbes Quadrangle, University of Pittsburgh, Pittsburgh, PA 15260 ($.60)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; *Criterion Referenced Tests; Test Construction; *Testing Problems; Test Validity
IDENTIFIERS Domain Referenced Tests

ABSTRACT

A framework, or classification scheme, is provided for displaying the spectrum of criterion-referenced tests. This framework illustrates that no single type of test can be identified as the absolute prototype criterion-referenced test It is shown that over the past 115 years criterion-referenced testing has grown to be a many-faceted concept, its multitude of specific instances differing qualitatively from each other. A definition is provided for a criterion-referenced test as one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Three broad categories are identified (well-defined domains, ill-defined domains, or those that are basically undefined), and the relationships are clarified. The classification scheme for distinguishing the varieties of these tests is presented in tabular form according to the basis for test development. Other tables give examples of criterion-referenced tests, and summarize additional breakdown of categories of tests which reference scores to unordered domains. Essentially, this framework is seen as a first step toward clarity of communication among professionals and toward improved test development.
(Author/GSK)

Working Paper #39

of the

Program in Educational Research

University of Pittsburgh, 1980

INDIVIDUAL DIFFERENCES BETWEEN CRITERION-REFERENCED TESTS*, **

by

Anthony J. Nitko

A paper presented at the 1980 Annual Meeting of the American Educational Research Association, Boston, Massachusetts, April 7-11, 1980.

Individual Differences Between Criterion-Referenced Tests*, **

In 1960, the term criterion-referenced testing was unknown to educators, psychologists, and measurement specialists. By 1978, however, over 600 references existed on the subject (Hambleton, Swaminathan, Algina, and Coulson, 1978) and the term was being used by many school people and most educational testing specialists. However, as can be testified by many who have used and worked with this type of testing, there exists considerable uncertainty as to what is meant by criterion-referencing. William Gray (1978), for example, did a content analysis of 57 descriptions of criterion-referencing written by nearly 40 authors and found that not only do different writers use the term differently, the same writers are sometimes inconsistent within the same article. This clearly demonstrates that there is no single agreed upon definition of criterion-referencing.

The purpose of this paper is to provide a framework or classification scheme that can be used to display the spectrum of criterion-referenced tests. This framework illustrates that no single type of test can be identified as <u>the</u> prototype criterion-referenced test. It illustrates that over the years criterion-referenced testing has emerged as a many-faceted concept, having a multitude of specific instances, that differ qualitatively from each other.

One reason for considering such a taxonomic classification of the area is that it is one step on the road to systemization. Systematizing a body of knowledge can lead to advancing the work in an area by pointing out what has been done and what yet remains to be done. Also, it permits the similarities and differences among the works of various investigators to be displayed even though their rhetorics have long since become amalgamated in common useage.

Basic Distinctions

A broad definition can be used to distinguish criterion-referenced tests from others (Glaser and Nitko, 1971, p. 653):

> <u>A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards.</u> Performance standards are generally specified by defining

---

* Helpful comments and suggestions of Robert Glaser, C. Mauritz Lindvall, and Ronald K. Hambleton are greatfully acknowledged.

** A nine-page bibliography of theoretical and how-to-do-it articles organized according to the classification scheme presented in this paper is available. Please send a self-addressed stamped envelope.

> a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain, and such measures are referenced directly to this domain for each individual measured. [Criterion-referenced tests] ... are specifically constructed to support generalizations about an individual's performance relative to a specific domain of tasks .... The term task includes both content and process.

Many types of achievement tests seem to fit this broad definition. The first step taken to distinguish among them is to characterize the manner in which each defines the domain of behaviors to which an examinee's test performance is to be referenced. Three broad categories can be identified: Domains are either well-defined, ill-defined, or basically undefined. Tests falling into the latter two categories do not qualify as criterion-referenced tests under the broad definition adopted here, even though the test developers may claim otherwise.

Examples of tests based on ill-defined domains include:

(1) Tests developed from "behavioral objectives" that are so poorly written and ambiguous that it is not possible to know to which domains of behavior a test score can be referenced, or
(2) Tests developed in such a way that the domain is defined only in terms of the particular items on the test, so that the broader generalizations, which are required for decision-making, cannot be made.

Tests based on ill-defined domains have been called "cloud-referenced tests" by Popham (1974).

Some tests referred to as criterion-referenced simply do not define a domain of behavior and, thus, such tests cannot form the basis for referencing test performance in the manner considered here. Frequently, when you encounter such tests, you will notice that the test developer has confused the idea of a cut-off score with the idea of criterion-referencing an examinee's score to a domain of instructionally relevant performance. These kinds of tests might be called pseudo-referenced tests and represent a misapplication of the idea of criterion-referencing as described here.

A domain is well-defined when it is clear to both the test developer and the test user which categories of performance (or which kinds of tasks) are and are not to be considered as potential test items. Well-defined domains are a necessary condition for criterion-referencing since the basic idea is to generalize how well a student can perform in a broader class of behaviors, only a few of which happen to appear on a particular test form. Since test development includes much more than defining a domain, the con-

dition of domain definition, while necessary, is not sufficient.

Insert Table 1 About Here

The Classification Scheme

Table 1 shows the scheme for classifying various kinds of criterion-referenced tests. Notice that the column headings identify the ways to characterize domains on which tests are built. In the scheme, well-defined domains are further sub-divided into ordered and unordered domains. This distinction is a fundamental one and is based on a conception that in some cases the behaviors in a domain can be ordered along a continuum of achievement such as that which Glaser and Klaus (1962; Glaser, 1963) and others have spoken. For example, one can think of ordering behavior in a sequence of learning prerequisites so that a test can be built which references a student's performance to this sequence. One use of such a test would be to provide information about what has been already learned and thus permits the planning of the next stage of instruction. Five different basis for ordering are listed in the table and perhaps there are more. These different bases will be described and illustrated shortly.

The scheme illustrated in Table 1 shows four broad categories for classifying criterion-referenced tests that are developed from well-defined but unordered domains. Perhaps other ways not included in the table could be listed, too.

Before describing and illustrating the various kinds of criterion-referenced tests that fall into each of these categories, it should be noted that the basis used to categorize them was not just the original authors' or test developers' definitions, but a broader consideration of (a) the tests themselves, (b) the manner in which they were produced, and (c) the overall context of the authors' discussions of them. As a result of this, you will notice that there are included several tests or suggestions which have not been identified previously as criterion-referenced. Some of these existed before Glaser and Klaus invented the term. Other tests and procedures are identified as criterion-referenced -- even though their authors explicitly deny that they are -- if they appear to satisfy the broad definition adopted. Frequently, authors deny their association with criterion-referencing because they disagree with one particular definition, interpretation, or application, but they fail to be specific about the nuance of the concept with which they disagree. One possible outcome of the present treatment is to display the basis for many of these disagreements and to provide a framework into which a wide range of suggestions for test improvement may have pluriaxial existence. Still another outcome of this process is to identify distinctions between various types of criterion-referencing that the original authors themselves have either not recognized or have not made explicit.

Insert Table 2 About Here

Table 1. A scheme for classifying and distinguishing the many varieties of tests that have been called criterion-referenced

Basis for test development	How the domain of behavior for achievement testing is characterized			
	Well-defined and ordered domains	Well-defined but unordered domains	Ill-defined domains	Un-defined domains
	Ordering based on judgements of the social or aesthetic quality of an examinee's product or performance.	Specifying the stimulus properties of the items to be included in the domain.	Poorly articulated behavioral objectives.	No attempt to define a domain to which to reference test performance.
	Ordering based on which level of difficulty or complexity a topic or subject is learned.	Specifying the stimuli and the responses in the domain.	Defining the domain only in terms of the particular items on the test.	Using a cut-off score but not defining a performance domain.
	Ordering based on degree of proficiency with which a complex skill is performed.	Specifying the "diagnostic" categories of the domain.		
	Ordering based on prerequisite sequences for acquiring an intellectual or psychomotor skill.	Specifying the abstractions, traits, or constructs that define the domain.		
	Ordering based on an empirically defined latent test.	Other ways of specifying the domain are possible.		
	Ordering on other bases are possible.			

Table 2

Various categories of criterion-referenced tests based on well-defined and ordered domains

Basis for scaling or ordering the defined domain of behavior[a]	Examples[b]
Judged social or aesthetic quality of the performance	Rev. George Fisher's Scale Books (1864)
	E. L. Thorndike's Handwriting (1909) and Drawing (1913) Scales
Complexity or difficulty level of the subject-matter	Ayre's Spelling Scale (1915)
	Glaser's Criterion-Referenced Measures I (1962, 1963)
	Cox and Graham's Arithmetic Scale (1966)
Degree of proficiency with which complex skills are performed	Harvard-Newton English Composition Scales (1914)
	Glaser's Criterion-Referenced Measures II (1962, 1963)
	Perhaps certain sports events or physical fitness tests
Prerequisite sequence for acquiring intellectual and psychomotor skills	Gagné's Learning Hierarchies (1962)
	Piagetian Development Scales (Gray, 1978)
	Infant Development Scales (Uzgiris & Hunt, 1966)
Location on an empirically defined latent trait	Connolly, Nachtman, and Prichett arithmetic tests (1971)
	Other tests build with latent trait models (e.g., Rasch, 1960 or Birnbaum, 1968) provided they are referenced to well defined and ordered domains of behavior.

[a]Other bases for scaling are possible as well; [b]Examples are meant to be illustrative rather than representative or exhaustive.

Table 2 shows and gives references to examples of criterion-referenced tests* that are based on well-defined and ordered domains. Notice that the examples span roughly 100 years -- from Rev. George Fisher's 1864 Scale Books to some current conceptions of criterion-refereneing that use one of the latent trait theory models.

Among the distinctions made, is the one between the two types of orderings advocated by Glaser (1963; Glaser and Klaus, 1962). One ordering is based on subject-matter difficulty or complexity: an examinee's score is scaled to reveal to which level of difficulty or to which level of complexity a topic or subject has been learned. Figure 1 shows an example of this based on a simple addition scale developed by Cox and Graham (1966) to identify the most complex type of problem a child could perform. (The figure also illustrates the idea that norm-referencing is not incompatible with criterion-referencing.)

A second type of ordering advocated by Glaser is based on the degree of proficiency with which complex intellectual or psychomotor skills can be performed. A summary of the various ways proficient experts perform differently from novices has been provided by Chi and Glaser (in press).

Insert Figure 1 and Table 3 about here

Table 3 summarizes an additional break-down of the categories of tests which reference scores to unordered domains. There are four broad categories which distinguish the tests according to both the manner and specificity with which the domains are defined. Within each category there are certain nuances. These within category nuances have in common essentially the same basis for defining and delineating a domain, but each emphasizes a somewhat different perspective or aspect. The time constraints of this presentation do not permit a full discussion of each category, but some comments can be made.

With regard to the first category -- stimulus properties and sampling plans -- it should be noted that for purposes of test development, it is necessary to use intuition or a "theory of performance" to specify those stimulus properties that would likely cause behavior to vary and, hence, that ought to be taken into account when sampling from the domain. Thus, while focus is on stimulus characteristics, response characteristics are not neglected. When a theory of performance is crude or undeveloped, stratification and sampling follow suit.

It should be noted, too, that Ebel's (1962) content-standard scores are placed in this first category. Although Glaser (1963) has pointed to the similarity between his proposal and Ebel's content-standard scores, Ebel

---

* In the psychometric literature such ordering of performance is often referred to as scaling. Scaling implies establishing a metric as well as determining ordinality (cf. Angoff, 1971).

Table 3

Various categories of criterion-referenced tests based on well-defined but unordered domains

Basis for delineating the behavior domain[a]	During test development emphasis is placed on:	Examples[b]
Stimulus Properties of the Domain and the Sampling Plan of the Test	Defining content and content strata	Starch's English Vocabulary Test (1916)
		Ebel's Content-standard English Vocabulary Test (1962)
	Specifying stimulus properties of item domains	Hively's Item Forms (1966, 1968)
		Osburn's Item Forms (1968)
	Specifying the precise relationship between instructional content and item domain	Bormuth's transformational Rules (1970)
Verbal Statements of Stimuli and Responses in Domain	Behavioral objectives with or without the cut-off score ("criterion") specified	Tests based on Mager's Type of Objectives (1962)
		Curriculum Embedded Tests of IPI Mathematics (1967)
		Popham and Husek's Criterion-Referenced Testing (1969)
		Harris and Stewart's Criterion-Referenced Testing (1971)
	Elaborated descriptions of behaviors and stimuli	Popham's Criterion-Referenced Tests (1975, 1978)
		IOX Objectives-Based Tests (1972): Amplified Objectives
		IOX Test Specifications (1978)

[a]Other bases for delineating exist; [b]Examples are meant to be illustrative rather than representative or exhaustive.

Table 3 (continued)

Basis for delineating the behavior domain[a]	During test development emphasis is placed on:	Examples[b]
"Diagnostic" Categories of Performance	Identifying entry level behaviors	Hunt and Kirk Tests of School Readiness (1974)
	Identifying behavior	Tests build on Resnick's Component Analysis (1973)
		Gagné's Two Stage Testing (1970)
	Identifying and categorizing erroneous responses	"Tab-Item" Technique (1954)
		Nesbit's CHILD Program (1966)
		Hsu's Computer - Assisted Diagnostic Tests (1972)
	Identifying erroneous processes	Beck's Blending Algorithm (1972)
		Interviews to determine what processes were used in responding
Abstractions, Traits or Constructs	Specifying specific behaviors or categories of behaviors that delimit the abstraction, trait, or construct	Tests based on the Taxonomy of Educational Objectives (1956)
		Certain basic skills survey tests, e.g., ITBS, MAT

[a]Other bases for delineating exist; [b]Examples are meant to be illustrative rather than representative or exhaustive.

has argued against criterion-referenced tests while still implying the usefulness of content-standard scores (Ebel, 1970; 1971; 1978). It is not clear whether this debate centers on the whole of criterion-referenced testing or only on certain varieties of these tests.

It is the second category -- delineating a domain by specifying stimuli and responses -- that has received the most publicity, professional attention, and practical work. Further, most discussions of criterion-referenced testing have assumed that it is necessary to use some variety of behavioral objectives in order to develop a criterion-referenced test.

As is indicated by the third category, "diagnosis" has taken a variety of forms including identifying such aspects as (a) entry level behaviors, (b) missing component behaviors, (c) erroneous responses, and (d) erroneous processes.

Perhaps the most controversial category is the fourth. Tests in this category specify the domains in terms of abstractions, traits, or constructs and frequently use fine-grained, behavioral objectives as well. The categories of the Bloom, et al. (1956) Taxonomy, for example, refer mainly to internal processes or psychological constructs (e.g., see Cronbach, 1971). Reading comprehension or spelling ability are other examples of constructs or traits.

It may be thought that these tests are really "cloud-referenced" -- and indeed there are some tests for which this seems true. But the distinction here is that if the tests do have reasonably well-defined domains of instructionally relevant behaviors, they fall within the scope of the broad definition of criterion-referencing adopted here. If the developers of such tests choose to define these domains in terms of abstractions or constructs, rather than narrow stimulus/response classifications, perhaps this may diminish the usefulness of the tests for certain purposes in particular instructional programs. Nevertheless, for many such tests, the descriptions of the domains are understood by most teachers and educators, and, in that sense, can be considered to be well-defined.

Implications

Among the implications of such a classification scheme for criterion-referenced tests are these.

1. It is generally recognized by the profession that there are many poorly constructed criterion-referenced tests and some attempt has been made develop sets of guidelines or standards for evaluating them (e.g., Hambleton, and Eignor, 1978). Unless care is taken, however, such guidelines are likely to focus on only one or two varieties of criterion-referencing. The result may be that many useful tests are judged unfairly. Further, the concept of criterion-referencing that is communicated through such standards, to the profession and to users of tests, is likely to be woefully incomplete unless a broad view is taken.

2. Traditional concepts of reliability and validity appear much more relevant to the total field of criterion-referencing than has been previously

admitted. Looming over the whole area, for example, is the notion of construct validation. Among the many interpretations that are advocated for different kinds of criterion-referenced tests are such ideas as: mastery vs. non-mastery, expert vs. novice, hierarchical learning sequence, and a host of diagnostic categories each with specific implications. Criterion-oriented validity is another traditional concept that appears applicable to many of the tests in the field which lay claim to being useful for labeling or diagnosing children or for assigning them to various, qualitatively different instructional treatments.

3. A third implication is that in order for professional work in this area to continue it is necessary to make careful distinctions among various types of tests and to avoid intermingling substantively different rhetorical arguments. One reason why there are many definitions and varieties of criterion-referenced tests is because of the changing nature of the concept as well as how it has been shaped by various applications. This is to be expected in any area that is growing and maturing and need not be disconcerting if reasonable care is taken to properly reference one's definitional source.

4. Criterion-referenced testing started as a movement to make tests more related to the kinds of information needed for effective instructional decisions (Glaser, 1963). One can ask which, if any, of these many types of criterion-referenced tests are able to fulfill that original intention and which types are compromising it. Further critical analysis is needed to identify the potential of each type for the improvement of the learning environment and further work should be done where necessary.

5. Two very popular definitions, Glaser's (1963; Glaser and Klaus, 1962) and Popham's (1975; Popham and Husek, 1969), appear as quite different in intention. Most of the psychometric work on criterion-referencing has been directed toward the Popham ideas which consider primarily domains or collections of essentially unordered behaviors or tasks. The testing problem is seen as one of estimating an examinee's status on a domain, usually in terms of a proportion of tasks that can be performed in the entire domain. On the other hand, much of the relevant work by psychologist in such areas as cognitive processes, novice/expert distinctions, and problem solving strategies has been more in line with the Glaser concept of an ordered domain. There appears, then, to be a continuing need for intercourse between the three kindgoms of Educationdom, Learningvania, and Psychometrica (Glaser, 1969)on this matter of what is important to test and how to go about testing it. Perhaps a scheme such as the one presented here will be a first step toward clarity of communication in this regard and in regard to some of the other measurement issues raised.

Summary

This paper provides a framework for integrating the many conceptions of criterion-referenced testing. The categories of the scheme that is developed are illustrated with examples of criterion-referenced tests which have been developed over a period of one and a quarter centuries: between 1864 and 1978. The scheme illustrates how various conceptions of criterion-referenced testing are both similar to and different from each other. This clarifies the relationships between some well-known and popular definitions. The framework is seen as a first step toward clarity of communication among professionals and toward improved test development.

Reference Notes

1. Uzgiris, I. C., & Hunt, J. McV. An instrument for assessing infant psychological development. Unpublished paper from the Psychological Development Laboratory, University of Illinois, 1966.

2. Uzgiris, I. C., & Hunt, J. McV. Oridnal scales of infant psychological development: Information concerning six demonstration filss. Unpublished paper from the Psychological Development Laboratory, University of Illinois, 1968.

3. Hunt, J., McV. Utility of ordinal scales derived from Piaget's observations. In I. C. Uzgiris (Organizer), Infant development from a Piagetian approach. Symposium presented at the meetings of the American Psychological Association, Montreal, Canada, August, 1973.

4. Hively, W. Preparation of a programmed course in algebra for secondary school teachers: A report to the National Science Foundation. Minnesota State Department of Education, Minnesota National Laboratory, 1966.

5. Popham, W. J. A lasso for runaway test items. A presentation at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, D. C., October, 1978.

6. Hsu, T. C., & Carlson, M. Computer-assisted testing. Unpublished manuscript, Univer ty of Pittsburgh, Learning Research and Development Center, 1972.

7. Binstock, L., Pingel, K., & Hsu, T. C. The design of a computer-assisted mastery testing model for diagnosing inaccurate use of processes underlying problem solutions. Unpublished manuscript, University of Pittsburgh, Learning Research and Development Center, November, 1975.

8. Harris, M. L., & Stewart, D. M. Application of classical strategies to criterion-referenced test construction: An example. In M. R. Quilling (Chair.), Criterion-referenced tests: Sense and nonsense. Symposium presented at the Annual Meeting of the American Educational Research Association, New York City, 1971.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Ayres, L. P. A measuring scale for ability in spelling. New York: Russell Sage Foundation, 1915.

Ballou, F. W. Scales for the measurement of English composition. Harvard-Newton Bulletins (No. 2). Cambridge, MA: Harvard University, 1914.

Beck, I. B. Comprehension during the acquisition of decoding skills. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh, 1977. (Publication No. 1977/4)

Beck, I. B., & Mitroff, D. D. The rationale and design of a primary grades reading system for an individualized classroom. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh, 1972. (Publication No. 1972/4)

Birnbaum, A. Chapters 17-20. In Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publisheng Co., 1968.

Bloom, B. S. (Ed.) Taxonomy of educational objectives. Handbook I. Cognitive Domain. New York: David McKay, 1956.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Chadwick, E. Statistics of educational results. The Museum, A Quarterly Magazine of Education, Literature and Science. 1864, 3, 479-484.

Chi, M., & Glaser, R. The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement In Proceedings of the CSE 1978 Conference: Issues in measurement and methodology. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, in press.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. Key Math Diagnostic Arithemtic Test. Circle Pines, MN: American Guidance Service, Inc., 1971.

Cox, R. C., & Boston, M. E. Diagnosis of pupil achievement in the Individually Prescribed Instruction Project. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh, 1967. (Working Paper No. 15)

Cronbach, R. C., & Graham, G. T. The development of a sequentially scaled achievement test. Journal of Educational Measurement, 1966, 3, 147-150.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 11-17.

Ebel, R. L. Some limitations of criterion-referenced measurements. In Testing in turmoil: A conference on problems and issues in educational measurement. Greenwich, CT: Educational Records Bureau, 1970.

Ebel, R. L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Ebel, R. L. The case for norm-referenced measurements. Educational Researcher, 1978, 7 (11), 3-5.

Farr, R. C., Prescott, G. A., Balow, I. H., & Hogan, T. P. Teacher's manual for administering and interpreting: Metropolitan Achievement Tests. New York: The Psychological Corporation, 1978.

Gagné, R. M. The acquisition of knowledge. Psychological Review, 1962, 69, 355-365.

Gagné, R. M. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Gagné, R. M. The conditions of learning (2nd ed.). New York: Holt, Rinehart & Winston, 1970. (a)

Gagné, R. M. Instructional variables and learning outcomes. In M. C. Wittrock & D. Wiley (Eds.), Evaluation of instruction. New York: Holt, Rinehart & Winston, 1970, (b)

Gagné, R. M., Major, J. R., Garstens, H. L., & Paradise, N. E. Factors in acquiring knowledge of a mathematical task. Psychological Monographs, 1962, 76, (7, Whole No. 526).

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75, (14, Whole No. 518).

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

Glaser, R., Damrin, D. E., & Gardner, F. M. The tab item: A technique for the measurement of proficiency in diagnostic problem solving tasks. Educational and Psychological Measurement, 1954, 14, 283-293.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. Gagné (Ed.), Psychological principles in systems development. New York: Holt, Rinehart & Winston, 1962.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

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

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

Hieronymus, A. N. Today's testing: What do we know how to do? In Proceedings of the 1971 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1972.

Hieronymus, A. N., & Lindquist, E. F. Teacher's guide for administration, interpretation, and use: Iowa Tests of Basic Skills. Boston: Houghton-Mifflin, 1971.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project. CSE Monograph Series in Evaluation, (No. 1) Los Angeles: Center for the Study of Evaluation, University of California, 1973.

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

Hunt, J. McV. Implications of sequential order and hierarchy in early psychological development. In B. X. Friedlander, G. M. Sterritt, & G. E. Kirk (Eds.), Exceptional infant (Vol. 3). New York: Brunner/Mazel, 1975.

Hunt, J. McV., & Kirk, G. E. Criterion-referenced tests of school readiness: A paradigm with illustrations. Genetic Psychology Monographs, 1974, 90, 143-182.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Mager, R. F. Preparing instructional objectives. Palo Alto, CA: Feardon Publishers, 1962.

Nesbit, M. Y. The CHILD program: Computer help in learning diagnosis of arithmetic scores. (Curriculum Bulletin 7-E-B.) Miami, FL: Dade County Board of Public Instruction, 1966.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Popham, W. J. Developing IOX Objectives-Based Tests: Procedure guidelines. (Technical Paper No. 8) Los Angeles: The Instructional Objectives Exchange, August, 1972.

Popham, W. J. An approaching peril: Cloud-referenced tests. Phi Delta Kappan, 1974, 56, 614-615.

Popham, W. J. Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975.

Popham, W. J. As always, provocative. Journal of Educational Measurement 1978, 15, 297-300. (a)

Popham, W. J. The case for criterion-referenced measurements. Educational Researcher, 1978, 7(11), 6-10. (b)

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

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

Rasch, G. Probabilistic models for some intelligence and educational tests. Copenhagen, Denmark: The Danish Institute for Educational Research, 1960.

Resnick, L. B. The science and art of curriculum design. A paper prepared for the Career Education Task Force, National Institute of Education, United States Department of Health, Education and Welfare, 1974. (Also, Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh, Publication No. 1975/9.)

Resnick, L. B. Task Analysis in instructional design: Some cases from mathematics. In D. Klahr (Ed.). Cognition and instruction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1976.

Resnick, L. B., & Beck, I. L. Designing instruction in reading: Interaction of theory and practice. In J. T. Guthrie (Ed.), Aspects of reading acquisition. Baltimore, MD: Johns Hopkins University Press, 1976.

Resnick, L. B., Wang, M. C., & Kaplan, J. Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. Journal of Applied Behavior Analysis, 1973, 6, 679-710.

Starch, D. Educational measurements. New York: The Macmillan Co., 1916.

Thorndike, E. L. Handwriting. Teacher's College Record, 1910, 11, 1-93.

Thorndike, E. L. A scale for measuring achievement in drawing. Teacher's College Record, 1913, 14, 345-382.