DOCUMENT RESUME

ED 047 010 TM 000 384

AUTHOR	Nitko, Anthony J.
TITLE	Criterion-Referenced Testing in the Context of Instruction.
INSTITUTION	Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
PUB DATE	Oct 70
NOTE	19p.; Paper presented at the Educational Records Bureau-National Council on Measurement in Education Symposium, "Criterion-Referenced Measures: Pros and Cons," New York, New York, October 1970
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Academic Achievement, *Criterion Referenced Tests, Decision Making, Educational Objectives, Feedback, Individual Differences, *Instruction, Item Sampling, *Norm Referenced Tests, Scores, Standards, Student Behavior, Test Construction, *Testing, Test Interpretation

ABSTRACT

Criterion-referenced testing is defined and some of its background is discussed. A distinction is made between criterion-referenced scores, norm-referenced scores, cut-off scores, criterion scores, criterion variables, and content-standard scores. The relationship between norm-referenced information and criterion-referenced information is considered. The need for vigorous, empirically-based construct validation studies of criterion-referenced tests is pointed out. The use of criterion-referenced testing in instruction is considered in terms of absolute interpretations and mastery learning. It is seen that whether criterion-referenced testing and/or norm-referenced testing is needed to make instructional decisions depends upon the instructional context within which one operates. It is concluded that, for purposes of instruction and instructional decision-making, there is a need for the integration of measurement knowledge with knowledge about instructional psychology. (Author/TA)

Criterion-Referenced Testing in the Context of Instruction

Anthony J. Nitko
Learning Research and Development Center
University of Pittsburgh

A paper presented at the Educational Records Bureau-National Council on Measurement in Education Symposium, "Criterion-Referenced Measures: Pros and Cons," New York, New York, October 1970.

The preparation of this paper was supported by the Learning Research and Development Center supported as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare.

Abstract

Criterion-Referenced Testing in the Context of Instruction*

Anthony J. Nitko

University of Pittsburgh

Criterion-referenced testing is defined and some of its background is discussed. A distinction is made between criterion-referenced scores, norm-referenced scores, cut-off scores, criterion scores, criterion variables, and content-standard scores. The relationship between norm-referenced information and criterion-referenced information is considered. The need for vigorous, empirically-based construct validation studies of criterion-referenced tests is pointed out. The use of criterion-referenced testing in instruction is considered in terms of absolute interpretations and mastery learning. It is seen that whether criterion-referenced testing and/or norm-referenced testing is needed to make instructional decisions depends upon the instructional context within which one operates. It is concluded that, for purposes of instruction and instructional decision-making, there is a need for the integration of measurement knowledge with knowledge about instructional psychology.

*The preparation of this paper was supported by the Learning Research and Development Center supported as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare.

Criterion-Referenced Testing in the Context of Instruction[1]

When we talk about criterion-referenced testing, we need to distinguish it from some traditional usages of the word criterion with which it tends to be confused. The term criterion has been used many times in psychometrics to refer to a second variable which we are interested in predicting. For example, an aptitude test is sometimes said to predict a criterion such as end of course grades or scores on an achievement test. Sometimes the validity of a test is described in terms of its correlation with some criterion (or criteria).

A second common usage of criterion has been that of criterion scores. The criterion score functions much the same as a cut-off score for some decision. In this context, expressions such as "working to criterion level" have been employed. For example, a statement like: "this student answered 50 per cent of the test questions correctly, but has not reached the criterion level of performance which is answering 85 per cent of the questions correctly."

Neither of these two usages of the term criterion is quite what is meant by criterion-referenced testing. It is useful, therefore, to review some of the background for criterion-referenced testing in order to more clearly describe it.

---

[1]Grateful acknowledgement is made to Robert Glaser and Richard L. Ferguson for their helpful comments on the draft manuscript.

Criterion-Referenced Testing

Although it may be true that criterion-referenced tests were used earlier, the term can probably be attributed to Robert Glaser. It was first mentioned in connection with proficiency measurement in training (Glaser and Klaus, 1962) and later was applied to the measurement of educational achievement (Glaser, 1963). The motivation for this application to achievement measurement stemmed from a concern about the kind of achievement information required to make instructional decisions. Some instructional decisions concern individuals. For example, what kind of competence an individual needs in order for him to be successful in the next course in a sequence. Other decisions center around the adequacy of the instructional procedure itself. Tests which provided achievement information about an individual only in terms of how the individual compared with other members of the group, or which provided only sketchy information about the degree of competence he possessed with respect to some desired educational outcome, were not sufficient to make the kinds of decisions necessary for effective instructional design and guidance.

In his discussion, Glaser refers to two other people who had proposed similar ideas: John Flanagan (1951) and Robert Ebel (1962). Both the Flanagan and Ebel ideas, while similar to Glaser's, are different enough to warrant discussion.

The Flanagan reference is to his chapter on units, scores, and norms in Lindquist's (1951) Educational Measurement. Flanagan distinguished between five types of descriptive information that are necessary in order to interpret broadly educational achievement data. In that discussion he made a distinction between a "standard of performance" and a "norm-performance." A standard of performance on a test is defined as a desirable model or a

minimum goal we would like an individual to attain. A "norm-performance" is the present average performance or attainment with respect to a specific group or population. For example,

> The score of an individual as obtained on a French reading test might be at the tenth-grade norm. This gives little information about how well he reads various types of materials. The probable degree of comprehension of the individual in reading a typical French newspaper would provide a useful social standard for interpreting scores on a French reading test (pages 698-699).

He cautioned that it was unwise to use automatically and uncritically the present average test performance as the acceptable score for that test. The most fundamental piece of information that an achievement test should provide is a description of an individual's performance with respect to some defined body of content that can be interpreted without reference to the scores of other individuals or to norm groups.

Professor Ebel (1962) extended this distinction and presented two schemes for developing tests whose scores could be interpreted objectively and meaningfully without the use of norms. Of special emphasis are the content categories that the test items represent. One method would result in a display of selected test items along with descriptive information about how many of these items could be answered correctly by individuals at various total test score levels. For example, if 10 of the 50 mathematics items from the PSAT were displayed, it would be possible to make a statement like: "Persons with a standard score of 500 on the mathematics section of the PSAT will, on the average, get 4 or 5 of these 10 items correct." The selected items are obtained by first sorting a large number of items into subject-matter content categories, such as, calculations with fractions, verbal problems, triangles, circles, and so on. Then the one item in each category that best discriminates between the

high and low scoring groups on the entire test is selected to represent the content category. Data for assigning meaning to a score of 500 is obtained by finding how many of the ten items were answered correctly, on the average (the mode in this case), by those persons who had standard scores of 500. This is repeated for each standard score level.

A second more basic procedure for obtaining meaningful scores is to make the process by which the test is constructed systematic and explicit. This calls for a systematic sampling of test items, rather than a subjectively chosen collection of tasks. For, "unless the score is based on a systematic sample from a defined domain of tasks, it cannot provide a very sound basis for inferences as to the examinees' performance on similar collections of tasks (page 16)." As an illustration, tests were built that required the examinee to match definitions with words.

> "The tests were based on a spaced sample of 100 words from a specified dictionary. Explicit instructions were given [to the test constructors] for choosing a unique but representative sample, and for limiting the sample to words appropriate for the test. For each word the first synonym or defining phrase was copied from the dictionary . . . . These tests constitute one operational definition of the proportion of words in a certain dictionary for which a person 'knows' the meaning, and hence the size of his vocabulary in a certain sense (pages 24-25)."

The term "content-standard scores" was used to refer to the kind of scores derived from these tests. "Content" means that the score is based directly on the items comprising the test. "Standard" means both the common scale on which the scores are reported (per cent in this case) and the fact that the process by which the test is constructed, administered, and scored is made explicit and objective. Thus, an individual's obtained score is referred directly to the domain of content for interpretation. This is contrasted to normative-standard scores which are interpreted by referring to the performance of other individuals. It

should be noted that this is a different use of the word standard than was used by Flanagan, who used it in the sense of a minimum goal or a desired model.

In a way, Glaser (1962) combined both the notion of a desired model and the notion of a standard domain of content. He called for the specification of the type of behavior the individual is required to demonstrate with respect to the content. "The standard [or criterion] against which a student's performance is compared . . . is the behavior which defines each point along the achievement continuum (page 519)." A criterion-referenced test, then, is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those scores can demonstrate (Glaser and Nitko, 1970).

As an illustration, consider the problem of assessing the competency of a student in elementary school geometry. Competency in elementary geometry can be analyzed into a number of behavior classes. A test can be constructed to measure these behaviors and to give scores that can be interpreted in terms of them. On such a test, a score of 30 might mean that along with a number of lower level behaviors, the student is able to

> identify pictures of open continuous curves, lines, line segments, and rays; can state how these are related to each other; and can write symbolic names for specific illustrations of them. He can identify pictures of intersecting and non-intersecting lines and can name the point of intersection.

This score would also mean that the student could not demonstrate higher level behaviors such as

> identifying pictures that show angles; naming angles with three points; identifying the vertex of a triangle and an angle; identifying perpendicular lines; use a compass for bisection or drawing perpendiculars; and so on.

In like manner, a score of 20 might mean that the student could not demonstrate any of the behaviors implied by the higher scores, but could demonstrate all lower level behaviors, up to and including behaviors such as:

> naming the plane figures that comprise the faces of cubes, cones, pyramids, cylinders, and prisms; naming these solids; and identifying pictures of these solids.

It is apparent, then, that there are four characteristics inherent in criterion-referenced tests:

(1) the classes of behaviors that define different achievement levels are specified as clearly as is possible before the test is constructed.

(2) each behavior class is defined by a set of test situations (that is, test items or test tasks) in which the behaviors can be displayed in terms of all their important nuances.

(3) given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test.

(4) the obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior.

Norm-Referenced Scores from Criterion-Referenced Tests

Norm-referenced testing is well known. When a test is constructed to yield scores that can be interpreted in such a way as to determine an examinee's relative location in a population or group of other examinees who took the same test, then we have a norm-referenced test. Scores derived for norm-referenced information are reported as percentiles, standard scores, grade-equivalents or age-equivalents. To obtain these scores, the mean, standard deviation, and sometimes the form of the distribution is pre-specified.

It should be obvious that criterion-referenced testing can yield norm-referenced information. Under certain circumstances both criterion-referenced information and norm-referenced information are needed to make

a broad interpretation of an individual's test performance. Flanagan, Ebel, and Glaser all point this out.

In most circumstance one or the other kind of information is of primary concern. The test constructor can choose to maximize either criterion-referenced information or norm-referenced information, but seldom can he maximize both. Since norm-referenced scores derive most of their meaning from distributions in which we can distinguish one individual from another, judicious selection of test items with the help of statistical analysis will maximize this distinction. Such statistical selection of items for criterion-referenced tests makes little sense, however. The classes or domains of tasks which define a behavior are determined, insofar as is possible, before the test is constructed and then representative samples are drawn for inclusion on any test. To screen out some items for inclusion on a particular test because they possess desirable statistical characteristics will change the definitions of the behavioral categories (cf. Osburn, 1968). The kind of information desired when criterion-referenced tests are used is the behaviors an individual does or does not possess and whether or not the test yields meaningful normative-standard scores is often of secondary importance.

The Need for a Data Base

When one proceeds to build a criterion-referenced test he needs to be just as rigorous as when constructing a norm-referenced test. Given that the classes of behavior have been defined, empirical evidence is needed to support any contentions that the classes of test tasks do indeed reflect the behavior or competence of interest. There is a need for knowledge about test construction to become integrated with psychological knowledge and theory.

More often than not, a single verbal statement of a behavior implies that an individual ought to be able to perform quite a large domain of tasks. This is particularly true of instructional objectives, where generalization and transfer are of primary importance. These domains of tasks need to be systematically examined and, if necessary, stratified so that representative sampling can take place.

Most useful instructional objectives which are employed in curriculum design appear to be formulated as constructs. This is true because (1) the behavior that is referred to is most often stated in terms of a class of responses to a class of stimuli and (2) all of these statements are often tied together with psychological interpretations such as the need for prerequisites and the relationships among the objectives in the sequence of instruction. Specifications of the instructional objectives which are needed for criterion-referenced tests tend to avoid broad trait construct statements such as "reading ability." Thus, the job of building tests that have representative tasks defining classes of behavior becomes more difficult as the behaviors become more complex. It is easier to build tests to measure decoding skills than to measure reading comprehension. The basis for inference about "reading sbility," for example, is observable performance on the specified domain of tasks into which reading ability can be analyzed, such as, reading certain types of passages aloud, identifying objects described in a text, rephrasing sentences in a certain way, carrying out written instructions, reacting emotionally to described events, and so on. It would seem, then, that criterion-referenced test builders need to conduct many of the same kinds of construct validation studies as have been recommended for psychological tests and other kinds of achievement tests (Cronbach and Meehl, 1955; Cronbach, 1969).

Absolute Interpretation of Test Scores

Recently, Cronbach (1969) has called attention to the need for absolute interpretations of test performance. Criterion-referenced testing implies this also. Absolute interpretation refers to making judgments about a person's score in terms of what his performance on the test is and what that performance represents with respect to a defined domain of test tasks. It is contrasted to comparative or relative interpretations, by which judgments about a person's score are based on the scores of other individuals in the population or group to which he has membership. It is clear that the testing movement has given little attention to absolute interpretations (Cronbach, 1969).

Absolute interpretations can be extremely dangerous, however, if they are used inappropriately. Tests for which the domain of items is vaguely defined, for which the behaviors ellicted are indeterminate, and for which a representative sampling plan has been unspecified, are poor bases upon which to interpret scores in an absolute sense. Failure to perform proper analysis before test construction often leads to assessing only those educational goals that are easily measured. Such abuses are probably common in many classroom test interpretations--and, perhaps, in much of what is currently passing for criterion-referenced testing! As Professor Ebel (1962; 1970) points out, such abuses are reminiscent of the criticisms of the percentage course grade and of objective testing early in this century.

These abuses then point more strongly toward the need for properly constructed criterion-referenced tests, based on well defined and instructionally meaningful behaviors, in situations where absolute interpretations tend to be made or where these interpretations need to be made. This means replacing much of the "art" of item writing with the technology of item

writing: behavioral and task analysis, task construction, and domain specification. Such work is certainly not easy, but neither does it seem impossible. A few notable suggestions along these lines have been provided by Gagné (1969), Hively (1966; Hively, Patterson, and Page, 1968) and Bormuth (1970).

Mastery

Criterion-referenced tests have been employed most often in instructional situations where the notion of mastery learning is advocated. One issue in which criterion-referenced testing has become entangled is that of determining mastery. Some propose that a cut-off or "criterion score" needs to be established and that each student must be taught until he obtains a score greater than or equal to this cut-off score. Some have argued that the cut-off score must be located at the upper extreme since flawless performance is desirable.

Nothing about criterion-referenced testing implies any of this. That criterion-referenced testing does not depend on a cut-off score has been mentioned previously. Further, criterion-referenced testing does not imply a value judgment about whether flawless performance is desirable. It only seeks to assess what the behavior is.

Whether using cut-off scores with tests is good or bad, is an empirical question although it is embedded in the ethical and decision network within which one operates. For example, given that certain terminal outcomes are desired and that an instructional sequence is specified, the question is, what level of performance is required at each point in the learning sequence in order to maximize success at the next point in the sequence and so on until the terminal learning is attained. This appears to be a transfer of learning problem and not one which is left

entirely to subjective judgment. It is clear that such decisions cannot be based on poor information, such as a poorly constructed test, but must be based on the empirical findings of instructional psychology.

Related to criterion-referenced testing and mastery learning is the question of whether everyone needs to learn the same thing to the same degree and who imposses standards of competency. A reasonable discussion of this issue and its ethical implications is beyond the scope of the presentation. (For a cogent discussion of this issue in another context see Bandura (1969). Much of that discussion seems to apply to instruction.) Nothing in the nature of criterion-referenced testing implies that anyone necessarily meet a given standard of competency, only that such levels of competency be defined in terms of performance.

A humanistic point of view would take into account the goals of the individual as related to the goals of society and allow the individual to participate in choosing and planning his learning experiences. If the individual desires to become a "master" and is motivated to achieve mastery, then of necessity we must provide him with the experiences which will facilitate his becoming a master and provide him with assessments so that he can evaluate his progress toward the goal he has chosen. To be sure, this point has been made by others. An interesting recent example of the successful application of behavioral analysis is that given by Zoellner (1969) with respect to the teaching of English composition. He states the problem in this way:

> ". . . the central failure of current compositional pedagogy . . . is its apparent inability to furnish the student-writer with anything but the most generalized specification for getting from one side of the writing situation (poor writing) to the other (good writing). What is urgently needed is a pedagogical technique which will supply the student-writer with a set of compositional specifications which are a) successively intermediate rather than ultimate, b) visible rather than invisible, c) uniquely adapted to the student's unique writing problem, and d) behavioral rather than historical, addressed to writing rather than the written word (page 274)."

The Need for Norm-Referenced Information

So far this discussion has emphasized criterion-referenced information. The need for norm-referenced information as well as criterion-referenced information should be apparent. It is useful under certain circumstances to know not only what level of competency an individual or group has or does not have, but also how that competency is related to other individuals or groups which are similar in composition, have similar educational experiences, or which have similar aspirations. It is also important to know relative standing in groups that are basically different.

But "useful" can only be interpreted in terms of purpose. In order to determine what kind of information to collect or to emphasize, one needs to know what kind of decision needs to be made. In some decision contexts norm-referenced information is inescapable. It has been pointed out that in some parts of the world it may be that it is financially impossible to offer advanced education to all individuals. Here relative competency and relative standing with respect to all such applicants for education becomes one of the most important types of information that is needed for decision-making. Whether such a stance is valid is beyond the scope of this presentation. The answer to such a question, however, will determine to a large extent the type of information the educational decision-maker will need and the kinds of observations and data that will have to be collected.

Criterion-Referenced Testing vs. Norm-Referenced Testing

Is criterion-referenced information better than norm-referenced information? One cannot discuss the usefulness of one measurement procedure over another without knowing the context within which that information is

needed and how it will be used. As Green (1969) has noted, considerations of measurement per se are wasteful in the overall decision-making process. Failing to consider the interrelationship between measurement and decision-making neglects the importance of deciding what additional data need to be collected before adequate decisions can be made.

There is a difference between taking measurement for scientific purposes and testing in instructional situations. The scientist is concerned with the identification and measurement of stable properties and variables. He seeks to determine general laws and rules for determining the relationships between these variables. He is discipline oriented and this dictates to a large extent the variables he chooses to measure and the way in which he measures them. In the practice of instruction one is concerned primarily about what each pupil desires to learn and how to maximize the learning he desires. What is learned is of primary importance and is usually defined in terms of acquired behavior and competence. Instruction provides the conditions by which this learning takes place. In a somewhat different context Lord (1968) speaks to this point.

> It should be clear that there are important differences between testing for instructional purposes and testing for measurement purposes. The virtue of an instructional test lies ultimately in its effectiveness in changing the examinee. At the end, we would like him to be able to answer every test item correctly. A measurement instrument, on the other hand, should not alter the trait being measured. Moreover, . . ., measurement is most effective when the examinee knows the answers to only about half the test items. (page 2)

It is a platitudinous assertion that an educational system should provide for individual differences and should allow students at every level of ability to develop and excel. Several patterns of instructional procedures for adapting to individual differences as they appear in the school can be identified (Cronbach, 1967). One pattern occurs where

educational goals and instructional methods are relatively fixed and inflexible. Individual differences are taken into account by dropping students along the way. The underlying rationale involved is that every child should "go as far as his abilities warrant." A second pattern of adaptation to individual differences is one in which the prospective future role of a student is determined, and depending upon this role, he is provided with an appropriate curriculum. For example, vocationally oriented students get one kind of mathematics and academically oriented students get a different kind of mathematics. Generally in this type of adaptation individual differences the educational system has optional educational objectives, but within each option the instructional procedures are relatively fixed. A third pattern of adaptation to individual differences is one in which instructional procedures are varied to accommodate the differences in each student. Different students are taught differently and the sequence of what is learned is not common to all students. One way in which this pattern is implemented is to provide a fixed mainstream instructional sequence and to branch students to remedial work when needed. Upon completion of remedial work the student is returned to the mainstream instruction. Another way of implementing this pattern is to begin with an assessment of a pupil's learning habits and attitudes, achievements and skills, cognitive style, etc. This information is used to guide the student through a course of instruction that is uniquely tailored to his goals. Thus, students would learn in different ways and attain different goals.

Each of these different patterns of instruction will require different kinds of measurement that result from different types of information requirements and instructional decision-making requirements. It

is impossible then to apeak of the strengths and weakness of criterion-referenced or norm-referenced testing in a vacuum. The merits of any testing program lies in the extent to which it provides useful information to the decision-maker be he instructional designer, pupil, teacher, administrator, or the public at large.

Not only must this information be useful, but it must be usable as well. That is, the testing program must be designed into the instructional process so that the information that is required is easily obtained and available in a usable form at the time a decision needs to be made. Built into such an instructional system must be a procedure for conatantly updating and redefining the adequacy of the decisions being made and the information upon which they are based.

When viewed in this way, the distinction between testing and instruction becomes leaa diatinct, ao that the learner can look toward testing for feedback concerning his accomplishments and for guidance toward his chosen goals.

References

Bandura, Albert Principles of behavior modification. New York: Holt, Rinehart, and Winston, 1969.

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

Cronbach, Lee J. How can instruction be adapted to individual differences? In R. Gagné (ed.), Learning and individual differences. Columbus, Ohio: Charles E. Merrill Books, 1967, Pp. 23-29.

Cronbach, Lee J. Validation of educational measures. In Proceedingsof the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1969.

Cronbach, Lee J. and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 52, 1955, 281-302.

Ebel, Robert L. Content-standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, Robert L. Some limitations of criterion-referenced measurement. Symposium address at the American Educational Research Association, Minneapolis, March, 1970.

Flanagan, John C. Units, scores, and norms. In E. F. Lindquist (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1951, Pp. 695-763.

Gagné, Robert M. Instructional variables and learning outcomes. In W. C. Wittrock and D. Wiley (eds.), Evaluation of instruction. New York: Holt, Rinehart, and Winston, 1969.

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

Glaser, Robert and Klaus, David J. Proficiency measurement: Assessing human performance. In R. Gagné (ed.), Psychological principles in systems development. New York: Holt, Rinehart, and Winston, 1962, Pp. 419-474.

Glaser, Robert and Nitko, Anthony J. Measurement in learning and instruction. In R. L. Thorndike (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1970. (in press)

Green, Burt F. Comments on tailored testing. In W. Holtzman (ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1969.

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

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

Lindquist, E. F. (ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1951.

Lord, Fredrick M. Some test theory for tailored testing. Office of Naval Research Report. Princeton, New Jersey: Educational Testing Service, September, 1968.

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

Zoellner, Robert Talk-write: A behavioral pedagogy for composition. College English, 1969, 30.