DOCUMENT RESUME

ED 324 327 TM 015 485

AUTHOR Klein, Thomas W.
TITLE Characteristics Which Differentiate Criterion-Referenced from Norm-Referenced Tests.
INSTITUTION Nevada State Dept. of Education, Carson City. Planning, Research and Evaluation Branch.
PUB DATE Jul 90
NOTE 13p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Criterion Referenced Tests; *Differences; *Educational Assessment; Knowledge Level; Mastery Tests; Minimum Competency Testing; *Norm Referenced Tests; Outcomes of Education; Test Format; Test Items; *Test Use

ABSTRACT

Characteristics that distinguish criterion-referenced tests from their norm-referenced counterparts are discussed, including: the purposes that they are designed to serve; the characteristics of the types of items that they contain; and the manner in which they are developed. More specifically, the distinguishing characteristics include: reference for measurement; information obtained; homogeneity of item difficulty; standards of performance; distribution of outcomes; and item format. A critical difference is the reference for measurement. For criterion-referenced tests, the reference of measurement is the existence of the behavior of interest. The criterion-referenced test assesses whether or not the individual possesses the skill. The qualitative distinction about performance is optimal for assessing specific or minimum competencies. The reference for measurement for norm-referenced tests is the distribution of outcomes in a representative normative sample. The question asked is how the individual performs relative to others from the reference population. Performance standards are arbitrarily imposed on the norm-referenced measure. Normative tests are optimal when the entire range of knowledge or skill is of interest. These differences are summarized in a table. (SLD)

Characteristics Which Differentiate Criterion-Referenced from Norm-Referenced Tests

Prepared by:

Thomas W. Klein, Ph.D.
Program Director
Nevada Proficiency Examination Program

EUGENE T. PASLOV
Superintendent of Public Instruction

MARCIA R. BANDERA
Deputy Superintendent

MARTY SAMPLE
Deputy Superintendent

Planning, Research, and Evaluation Branch
KEVIN CROWE, *Director*

Capitol Complex, Carson City, Nevada 89710

Characteristics Which Differentiate Criterion-Referenced from Norm-Referenced Tests

Prepared by:

Thomas W. Klein, Ph.D.
Program Director
Nevada Proficiency Examination Program

Nevada Department of Education
Planning, Research and Evaluation Branch

July, 1990

NEVADA STATE BOARD OF EDUCATION

June M. Herrmann, President

Carley Sullivan, Vice President

Frank Brown, Member

Janice A. Clarke, Member

John K. Hill, Member

Liliam L. Hickey, Member

Kenneth W. Koester, Member

Marianne Long, Member

Yvonne Shaw, Member

Leanne Lawrence, Student Representative

Dr. Eugene T. Paslov
Superintendent of Public Instruction

Marcia R. Bandera, Deputy Superintendent
Instructional, Research and Evaluative Services

An Equal Opportunity Agency

Table of Contents

List of Tables

Introduction

Most educators, as well as a significant proportion of the general public, have at least an elementary knowledge of norm-referenced tests. Anyone who has been a student in this nation's elementary or secondary schools during the past 20 years could be expected to have taken at least one such test and had the results reported to his/her parents. Thus, the meaning of percentile rank and the significance of a norm group are widely understood. The scope of the application of norm-referenced measures in education, from the Child Language Ability Measure, which can be used with children as young as two years, through standardized tests of academic achievement, to the SAT and ACT college admission tests, make their avoidance almost impossible.

In contrast, criterion-referenced tests (by that name) are less familiar to many educators. This is true, despite the fact that the typical classroom examination that is used to determine whether a student has attained sufficient knowledge to pass a course, or to merit a specific grade, is a familiar example of this type of instrument. Another common form of criterion-referenced test is the licensing examination. Such an examination is used to insure that individuals possess at least certain minimum knowledge or skills before they are allowed to engage in activities that affect the public welfare. The concern is with the candidate's ability to meet specified minimum standards rather than the degree of expertise that the candidate might possess.

On the surface, criterion-referenced and norm-referenced tests need not appear significantly different. That is particularly true when criterion-referenced tests are combined in a single instrument designed to assess several levels of mastery or competency simultaneously. However, the two types of assessment instruments do differ in a number of significant respects, including:

1.) The purposes that they are designed to serve,
2.) The characteristics of the types of items that they contain, and
3.) The manner in which they are developed.

Some differences merely reflect the degree to which certain characteristics of the instrument are emphasized. Other differences are qualitative in nature and provide the clearest distinctions between the two types of instruments.

The purpose of this communication is to discuss a number of characteristics which differentiate criterion-referenced measures from their norm-referenced counterparts, and to provide a convenient summary of those differences.

Distinguishing Characteristics

Reference for Measurement

As their name implies, the reference for measurement is a critical characteristic that distinguishes between criterion-referenced tests and their norm-referenced counterparts. Unlike physical measures such as length or weight, psychological and educational measures generally do not have a physical reference or a scale that has a clearly recognized absolute zero. These measures must rely on scales which use as a reference either a well defined behavior or class of behaviors or the range of characteristics displayed by members of a specified population. Tests which have a behavioral reference, i.e. tests that require an individual to demonstrate the presence or absence of a trait or skill, are considered to be criterion-referenced measures. The test content is a direct representation of the criterion being assessed. Instruments which place an individual on a continuum defined by the range of behaviors of a particular type that is expressed in a specified population are referred to as norm-referenced tests. The test content generally reflects a variety of components of the characteristic being measured, which, when combined, provide the operational definition of the trait or skill. For the criterion-referenced test the reference for measurement is the existence of the variable, or the defined level of the variable, while the norm-referenced test relies on the distribution of the variable within a specified population.

Information Obtained

The essential differences in the references for these two types of measures suggest essential differences in the types of information that they yield. The criterion-referenced test asks the question, "Is the criterion met?", and provides either a yes or no answer. Either the individual possesses the skill, or assessed level of the skill, or he/she does not. The answer is essentially qualitative in nature. In contrast, the norm-referenced tests asks, "How much of the skill does the individual possess, in relation to the degree to which the skill is demonstrated in a known population?" The answer provided by a norm-referenced test is in terms of the proportion of the reference population which demonstrates a level of skill at or below that demonstrated by the individual tested. The percentile rank is the most elementary expression of norm-referenced test results and provides a statement of the general form, "The skill level demonstrated by this individual exceeds that demonstrated by a specified percentage of the reference population." A variety of numerical scales are generally developed for use in reporting the results or conducting statistical analyses of the results of norm-referenced tests. However, they all have as their basis the distribution of the trait in the reference population. Thus, the norm-referenced test provides information that is essentially quantitative in nature.

Homogeneity of Item Difficulty

The two types of tests also vary markedly in the range of the variable that the items are designed to assess. When applied to the measurement of achievement, all items in a criterion-referenced test focus on a single level of difficulty. The intent is to obtain the clearest indicator of the presence or absence of the targeted level of achievement, the criterion. Extreme homogeneity is a desirable characteristic of the item pool that defines the criterion-referenced test. Criterion-referenced instruments which reflect a variety of achievement levels can be constructed by combining several criterion-referenced tests of the same characteristic or skill. The components of these multi-level criterion-referenced tests are each independently scored to provide a yes or no answer for each level of the variable being assessed. To the extent that the variable being measured is hierarchical in nature, individuals who pass at the higher levels would be expected to have passed at all lower levels.

The intent of the norm-referenced measure is to provide as sensitive a measure as possible over the entire range of the variable represented in the population. Thus, the items in the norm-referenced achievement test can be expected to span the entire range of difficulty levels. The purpose is to create an instrument that maximizes the measured variation among individuals in the reference population. Tests composed of items that are scored as either correct or incorrect will include a large proportion of items that are of intermediate difficulty. A smaller proportion of items will be devoted to the assessment of achievement at the extremes of the distribution, using either very easy or very difficult items. In practice, applications of norm-referenced tests of achievement are frequently concerned with the identification of individuals at the low end of the scale. A greater proportion of easier than difficult items are usually included in norm-referenced tests designed with this application in mind.

With regard to the ideal level of homogeneity of item difficulty represented in the item pool for the two types of tests, criterion-referenced and norm-referenced tests can be considered to occupy the two extremes of a continuum. The optimal criterion-referenced test maximizes item homogeneity while the norm referenced test maximizes item heterogeneity. In the extreme, a norm-referenced achievement test might be considered to be a multi-level criterion-referenced test, with each item representing a single measure of a specific criterion level. However, norm-referenced tests are not scored in this manner and the difficulty level of norm-referenced items are generally determined by the proportion of individuals in the reference population that pass the item. In contrast, the difficulty level of criterion-referenced test items are determined, *a priori*, by standards which are developed to guide the processes of item development and standard setting.

Standards of Performance

A major characteristic relating to differences in the manner in which standards of performance are established for the two types of tests involves the point in the test development process when the standards are set.

For criterion-referenced tests, the test items themselves define the criterion. Thus, standard setting involves the review of each item that will contribute to the test to determine its appropriateness for the intended application. This standard setting is completed prior to test administration.

The standards for passing a norm-referenced test are not inherent in the instrument. The quality of the instrument is judged, in part, on the basis of its ability to differentiate among individuals on the dimension being measured. If the instrument is judged to be a valid measure of the dimension, and it demonstrates adequate reliability, it is judged to be an acceptable measure. Standards of performance are defined only after the test has been administered and the distribution of scores in the reference population, the norm group, has been determined. Most norm-referenced tests contain no inherent standard beyond the range of behavior that it may be used to assess, which is determined by item content. *Ad hoc* standards of performance are established for each application of the test, depending on the proportion of the norm group that the performance of the individual who is to be considered successful is expected to surpass. The norm-referenced test merely provides a ruler. The length that will be considered satisfactory is external to the test.

Distribution of Outcomes

The distribution of outcomes from the ideal criterion-referenced achievement test would be bimodal. That is, the percentage of individuals that did not meet the criterion would answer none of the items correctly, while the proportion of students that met the criterion would answer all items correctly. Given that there is some random error associated with each item that makes up the test, individuals who do not meet the criterion can be expected to answer some of the items correctly, due to chance alone. Individuals who meet the criterion would be expected to fail a few items for the same reason. Thus, the standard for judging whether an individual has met the criterion rarely requires perfect performance. For tests of educational achievement the standard frequently employed requires that 80% of the items be answered correctly. Despite the various sources of unreliability which are reflected in the scores on a criterion-referenced achievement test, the bimodal nature of the distribution of outcomes may still be apparent when those data are presented in graphic form.

The distribution of results for the ideal norm-referenced achievement test would look quite different. These tests would be expected to produce the familiar bell-shaped normal distribution, with most individuals scoring near the mean and few scoring at the extremes. No one would be expected to achieve either a zero or a perfect score. Unlike criterion-referenced tests, random error which is reflected in norm-referenced scores only reinforces the expectation that the resulting distribution of scores will be normal. In fact, the expected distribution for the combination of a number of totally random events is the normal distribution. Thus, the observation of a normal distribution of results from a norm-referenced achievement test is consistent with the expectation that the test is accomplishing the intended task, but it adds no weight to that conclusion. However, a significant deviation from the normal distribution observed in norm-referenced test results would be reason to question the appropriateness of the measure.

Item Format

Both norm-referenced and criterion-referenced tests can be built using the same item formats. Multiple-choice questions seem to be the most popular for measures of academic achievement, due largely to their objectivity and efficiency. However, the item format selected for a particular test should be appropriate to the purpose of the measure.

The optimal criterion-referenced test used in the selection of individuals who will be required to perform a particular task would be a work sample or samples. The test items would parallel, to the extent possible, the actual work situation. In education, the closest parallel might be the direct assessment of writing. Rather than asking an individual about his/her knowledge of language, and inferring the ability to write from that knowledge, the individual is asked to provide a sample. Tests of reading comprehension also provide an illustration of a test that closely parallels the activity where the skill level is of interest. The individual is asked to read and demonstrate, usually through answering a number of questions, what he/she has learned through reading. Criterion-referenced tests will frequently use a variety of item formats in a single instrument in order to provide as complete an assessment of the criterion as possible.

The efficiency of a norm-referenced test, in terms of the time and resources it requires, is a major consideration. Not only is the cost important in determining the utility of a test for a particular application, but it is also a critical consideration for the quality of the norms that will be developed. Test developers are less likely to assess the large numbers of individuals required to produce stable norms when the cost of administration and/or scoring is high. The high levels of reliability that can be achieved with the multiple choice format, the number of items that can be presented

in a short amount of time, along with the efficiencies of scoring, make it the format of choice where large-scale application is required. In addition, relatively high levels of validity can also be achieved through the inclusion of items that test a variety of aspects of the skill being measured. In testing educational achievement, multiple choice tests are often validated through demonstrating their high correlation with other, more direct, indicators of the skill being assessed.

Summary

As indicated in the Table, criterion-referenced and norm-referenced tests differ in a variety of ways. A critical difference is the reference for measurement. The reference for the criterion-referenced test is the behavior of interest. The question that it addresses is, "Does the individual demonstrate the behavior?" In educational assessment, the criterion-referenced test provides direct information about the individual's skills and performance standards are designed into the criterion-referenced instrument's items. The criterion-referenced test yields a qualitative, yes-or-no answer with regard to the individual's performance. Criterion-referenced tests are optimal when one wishes to assess specific or minimum competencies.

In contrast, the reference for measurement for the norm-referenced test is the distribution of outcomes in a representative normative sample. The question addressed is, "How does the individual perform relative to others from the reference population?" Performance standards are arbitrarily imposed on the norm-referenced measure. The norm-referenced test tells what percentage of the population the performance exceeds, but may not be particularly informative about the nature of the skill or knowledge demonstrated. Norm-referenced tests are optimal when the entire range of knowledge or skill is of interest.

Summary of Characteristics that Differentiate Criterion-Referenced from Norm-Referenced Tests

Criterion-Referenced	Norm-Referenced
Reference for Measurement	
The criterion behavior of interest, determined by test content and difficulty level.	The distribution of the trait, within a range specified by test content, in an external reference group.
Information Obtained	
A qualitative (Yes/No) answer to the question, "Is the criterion met?"	A quantitative answer to the question, "Where does the performance of the individual rank relative to that of the reference group?"
Homogeneity of Item Difficulty	
Homogeneous relative to the criterion.	Heterogeneous for the broad range of achievement.
Standards of Performance	
Determined *a priori* and incorporated into the content of the test.	Not directly represented in the content of the test except for the range of content and the range of difficulty of the items.
Minimum Passing Score	
Usually 80% of the items representing the criterion.	Varies. Arbitrarily set, depending on the application.
Distribution of Outcomes	
Bimodal for presence or absence of criterion.	Continuous, often normally distributed.
Item Format	
Varied. Different formats often used in a single measure.	Varied. Multiple choice quite common. Often restricted to single format.
Major Applications	
Minimum competency, mastery, licensing.	Achievement testing, ranking of individuals on selection criteria.

Appendix 16

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 21,1991