DOCUMENT RESUME

ED 069 624 TM 001 108

AUTHOR Livingston, Samuel A.
TITLE A Classical Test-Theory Approach to Criterion-Referenced Tests.
PUB DATE 72
NOTE 12p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Ill., April, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; *Tests; *Theories

ABSTRACT

A criterion-referenced test is defined in this paper as any test for which the test user wants to compare each student's score not with the mean of some group, but with a specified criterion score, which does not depend on the scores the students actually obtain on the test. This definition, it is pointed out, implies that all the items on the test must measure the same thing. A classical test theory for criterion-referenced tests is derived.

AERA-NCME 1972

A CLASSICAL TEST-THEORY APPROACH TO CRITERION-REFERENCED TESTS

Samuel A. Livingston

Center for Social Organization of Schools

The Johns Hopkins University

The term "criterion-referenced test" has been defined in a number of different ways. The definition I propose is one that is broad enough to include all tests that would be included by most of the other definitions. According to this definition, a criterion-referenced test is any test for which the test user wants to compare each student's score not with the mean of some group, but with a specified criterion score, which does not depend on the scores the students actually obtain on the test. The mean score of the group being tested might fall far below the criterion score, or far above it, or anywhere in between. If the mean score of the group happened to fall exactly at the criterion score, then our criterion-referenced test would have all the statistical properties of a norm-referenced test--but of course this would be a coincidence that would not happen very often.

This definition implies that all the items on the test must measure the same thing--otherwise it makes no sense to specify a single criterion score. If the test user is interested in measuring specific behaviors, all the items on the test must measure the same specific behavior. What constitutes a specific behavior is a question that the test user must answer for himself. For one test user it might be "solving arithmetic computation problems;" for another it might be "solving multiplication problems; for another it might be "solving multiplication problems with a three-digit multiplicand and a two-digit multiplier, with zero in the tens position of the multiplicand;" and so on.

Once we have decided that we have a test made up of items that all measure the same thing, then we can apply the classical test theory model. We assume that each student's score on the test consists of two components: his true score, plus an error of measurement. Errors of measurement have an expected value of zero and do not correlate with true scores or with other errors of measurement. We also define the criterion score for the sum of two tests to be the sum of the two criterion scores.

Because the criterion-referenced test user is interested in comparing each student's score <u>not</u> with the group mean, but with the criterion score, we will have to revise some basic concepts. Variance, covariance, and correlation are all based on deviations from the mean, not from a criterion score--but we can easily define similar concepts based on deviations from the criterion score. These concepts are listed in Table 1. For want of any better terms, I call them the mean squared deviation, the mean product of deviations, and the criterion-referenced correlation coefficient. The standard error of measurement is listed here to emphasize the fact that it is <u>not</u> redefined; the standard error of measurement is the same for a criterion-referenced test as for a norm-referenced test.

Starting with these definitions and assumptions, we can derive classical test theory for criterion-referenced tests. Some of the results are listed in Table 2. As you might expect, both the derivations and the results are very similar to those for conventional norm-referenced classical test theory. Notice that the criterion-referenced reliability coefficient can be expressed as a ratio of mean squared deviations, just as the norm-referenced reliability coefficient can be expressed as a ratio of variances. Also notice that the Spearman-Brown formula, which relates reliability to test length, works exactly the same way for criterion-referenced tests as it does for norm-referenced.

Table 3 shows some formulas for expressing criterion-referenced indices in terms of norm-referenced indices. If a test user knows the norm-referenced reliability of his test, he can compute the criterion-referenced reliability. You can see from the formulas that the criterion-referenced reliability coefficient behaves the way we would expect a reliability coefficient to behave. As the variance of true scores increases, reliability increases. As the error variance increases, reliability decreases.

However, the criterion-referenced reliability coefficient also has some special properties of its own. Notice that a set of scores can have zero true variance and still have positive criterion-referenced reliability. In fact, a group of students randomly guessing at all the items would probably produce a set of scores with very high criterion-referenced reliability. In this case, the high reliability would mean that we can be quite confident that these students all have true scores below the criterion score.

Another important property of a criterion-referenced reliability coefficient is that its size depends on the distance between the criterion score and the mean score-- the $(\mu_x - C_x)$ term in the formula. The larger this difference--that is, the farther from the criterion score the group mean score happens to fall--the greater the reliability of the test. The reliability of the test is at its lowest when the group mean score happens to fall exactly at the criterion score. And when it does, the $(\mu_x - C_x)$ term becomes equal to zero, and the criterion-referenced reliability of the test equals its norm-referenced reliability. Therefore, the criterion-referenced reliability of a test can never be lower than its norm-referenced reliability and will usually be higher.

Figure 1 shows that the relationship between criterion-referenced reliability and norm-referenced reliability is linear, when the difference between the mean and the criterion score is held constant. Figure 2 shows how criterion-referenced reliability depends on this difference; here the norm-referenced reliability has been held constant. You can see that the curves in Figure 2 rise fastest at about one-half sigma. This is where a change in the mean or the criterion score will have the greatest effect on reliability--when the mean is about half a standard deviation from the criterion score.

So far, I have talked mainly about the reliability of criterion-referenced tests, but said very little about their validity. I believe that the validity of criterion-referenced tests can be established in the same way as the validity of norm-referenced tests. The technique I would favor is the multitrait-multimethod matrix suggested by Campbell and Fiske (1959). The only change we have to make is that when we are validating criterion-referenced tests, the correlations that we put into our matrix must be criterion-referenced correlations. This is an important difference, because it is possible for two tests which are uncorrelated, by conventional norm-referenced standards, to have a high criterion-referenced correlation. The reason is that a student who scores above the group mean on one test and below the group mean on the other test may have scored well above the criterion score on both tests.

Figure 3 illustrates this fact. In this example, test X and test Y are different criterion-referenced tests which are supposed to measure the same thing. They are not necessarily parallel tests; in fact, if this correlation is part of a validation study, they should not be parallel.

You can see that the norm-referenced correlation between X and Y is about zero; a student who scores high on one test, in relation to the group mean, is not especially likely to score high on the other test, in relation to the group mean. But a student who scores high in relation to the criterion score on one test is likely to score high in relation to the criterion score on the other test. Since most of the students in this group have scored high in relation to the criterion score on both tests, the two tests have a high positive criterion-referenced correlation.

It is also possible for two tests which correlate positively in the norm-referenced sense to have a criterion-referenced correlation which is negative. Figure 4 illustrates this situation. The criterion-referenced correlation is negative because many students who scored above criterion level on test X scored below criterion level on test Y.

What these examples show is that the criterion-referenced correlation between two tests depends heavily on the difficulty of the tests for the group of students being tested. Two tests can have a high criterion-referenced correlation only if they are of similar difficulty - a difficult math test would probably correlate more highly with a difficult reading test than with an easy math test. That may sound surprising to someone who is used to thinking in norm-referenced terms, but what it means is that if a student scores below criterion level on the difficult math test , he will probably score below criterion level on the difficult reading test, while he may be well above criterion level on the easy math test. Of course, the terms "difficult" and "easy" really refer to the students' level of proficiency in the skill that is being tested. A criterion-referenced test is "difficult" for a group of students if the criterion score represents a higher level of proficiency than most of the students have attained.

Why is it meaningful or useful to define a correlation that depends heavily on the choice of criterion scores? The reason is that when we use a criterion-referenced test, we are asking a question about the students' performance in relation to a particular criterion score. If we use the same set of items with a different criterion score, then we are asking a different question.

Table 1. Basic concepts.

Norm-referenced	Criterion-referenced
Mean: μ_x	Criterion score: C_x
Variance: $\sigma^2(X) = \mathcal{E}_i(X_i - \mu_x)^2$	Mean squared deviation: $D^2(X) = \mathcal{E}_i(X_i - C_x)^2$
Covariance: $\sigma(X,Y) = \mathcal{E}_i(X_i - \mu_x)(Y_i - \mu_y)$	Mean product of deviations: $D(X,Y) = \mathcal{E}_i(X_i - C_x)(Y_i - C_y)$
Correlation: $\rho(X,Y) = \dfrac{\sigma(X,Y)}{\sqrt{\sigma^2(X)\,\sigma^2(Y)}}$	Criterion-referenced correlation: $k(X,Y) = \dfrac{D(X,Y)}{\sqrt{D^2(X)\ D^2(Y)}}$
Standard error of measurement: $\sigma(E_x) = \sqrt{\mathcal{E}_j(X_j - T_x)}$	Standard error of measurement: $\sigma(E_x) = \sqrt{\mathcal{E}_j(X_j - T_x)}$

Note: $\mathcal{E}$ indicates the expected value.

Table 2. Important theorems.

	Norm-referenced	Criterion-referenced
Variance components	$\sigma^2(X) = \sigma^2(T_x) + \sigma^2(E_x)$	$D^2(X) = D^2(T_x) + \sigma^2(E_x)$
Covariance of true scores	$\sigma(X,Y) = \sigma(T_x,T_y)$	$D(X,Y) = D(T_x,T_y)$
Reliability	$\rho^2(X,T_x) = \rho(X,X') = \frac{\sigma^2(T_x)}{\sigma^2(X)}$	$k^2(X,T_x) = k(X,X') = \frac{D^2(T_x)}{D^2(X)}$
Spearman-Brown formula (where Y has n times as many items as X)	$\rho^2(Y,T_y) = \frac{n\,\rho^2(X,T_x)}{1 + (n - 1)\,\rho^2(X,T_x)}$	$k^2(Y,T_y) = \frac{n\,k^2(X,T_x)}{1 + (n - 1)\,k^2(X,T_x)}$
Correction for attenuation	$\rho(T_x,T_y) = \frac{\rho(X,Y)}{\sqrt{\rho^2(X,T_x)\,\rho^2(Y,T_y)}}$	$k(T_x,T_y) = \frac{k(X,Y)}{\sqrt{k^2(X,T_x)\,k^2(Y,T_y)}}$

Table 3. Conversion formulas.

Mean squared deviation:

$$D^2(X) = \sigma^2(X) + (\mu_x - C_x)^2$$

Mean product of deviations:

$$D(X,Y) = \sigma(X,Y) + (\mu_x - C_x)(\mu_y - C_y)$$

Criterion-referenced reliability coefficient:

$$k^2(X,T_x) = \frac{\rho^2(X,T_x)\sigma^2(X) + (\mu_x - C_x)^2}{\sigma^2(X) + (\mu_x - C_x)^2}$$

$$= \frac{\sigma^2(T_x) + (\mu_x - C_x)^2}{\sigma^2(X) + (\mu_x - C_x)^2}$$

$$= 1 - \frac{\sigma^2(E_x)}{\sigma^2(X) + (\mu_x - C_x)^2}$$

Values of the criterion-referenced reliability coefficient.

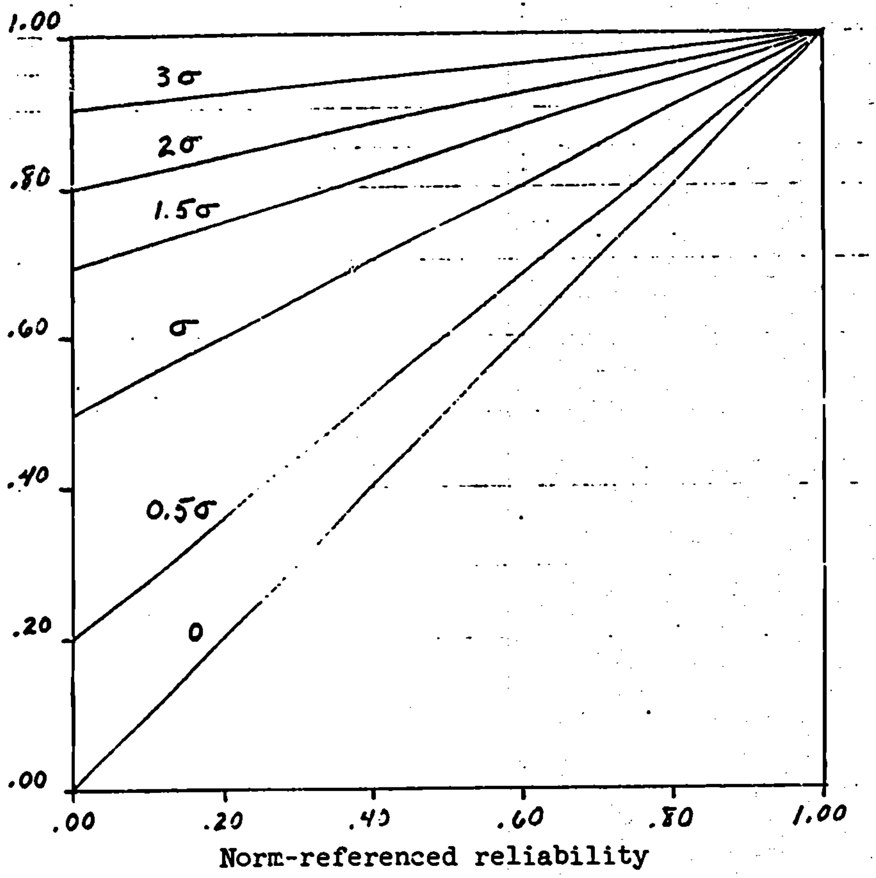

Figure 1.

Mean-criterion difference held constant at six specified levels.

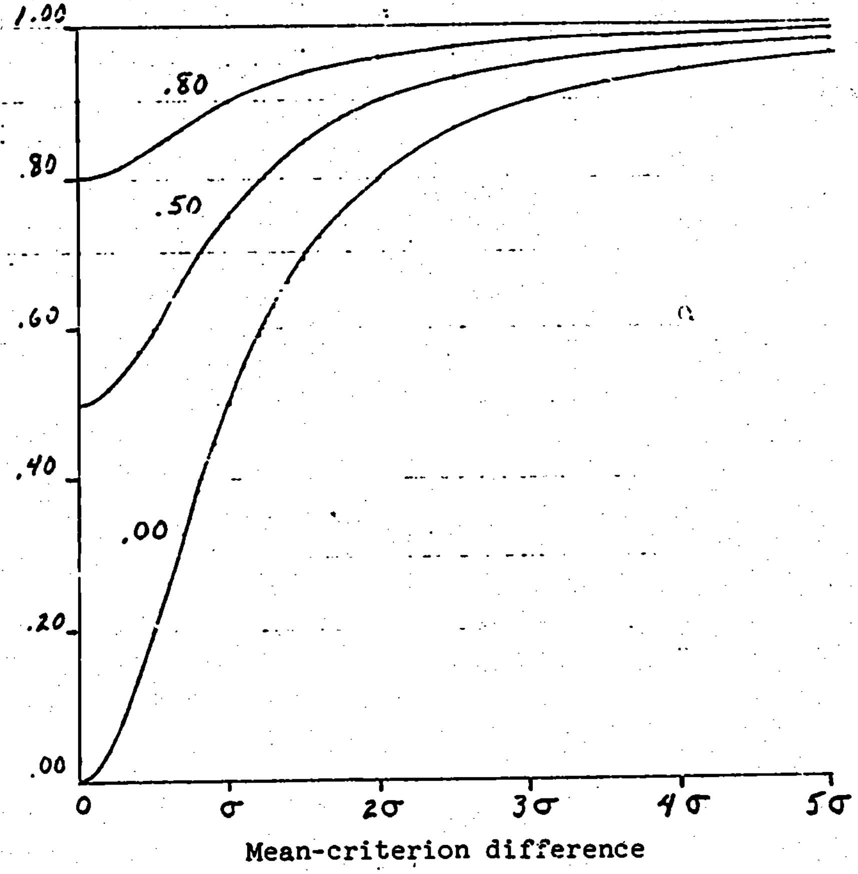

Figure 2.

Norm-referenced reliability held constant at three specified levels.

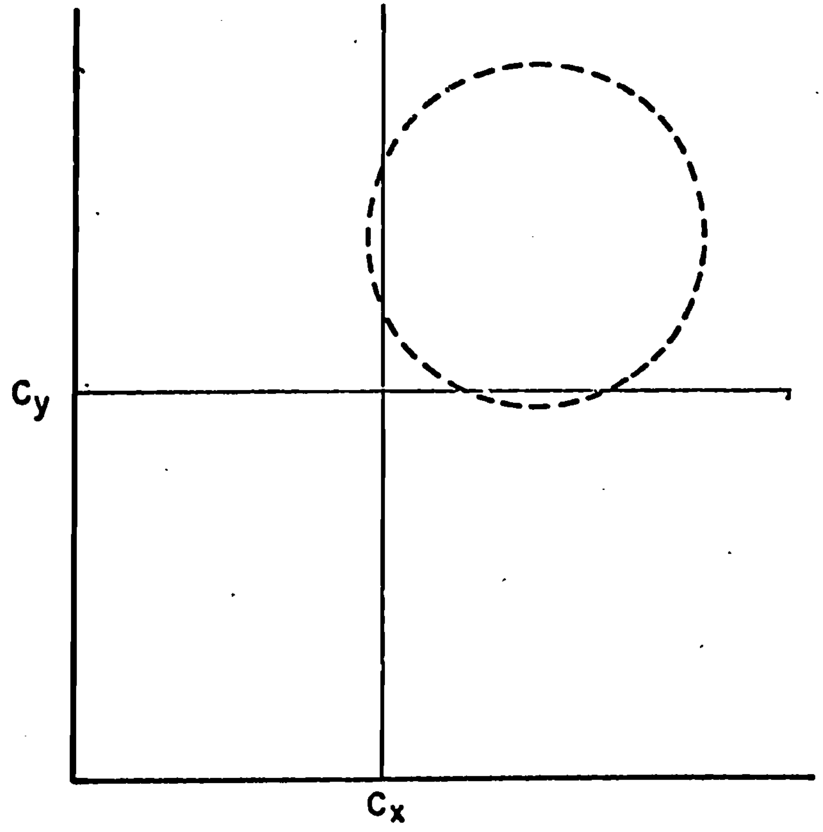

Figure 3. Norm-referenced Correlation is zero; criterion-referenced correlation is positive.

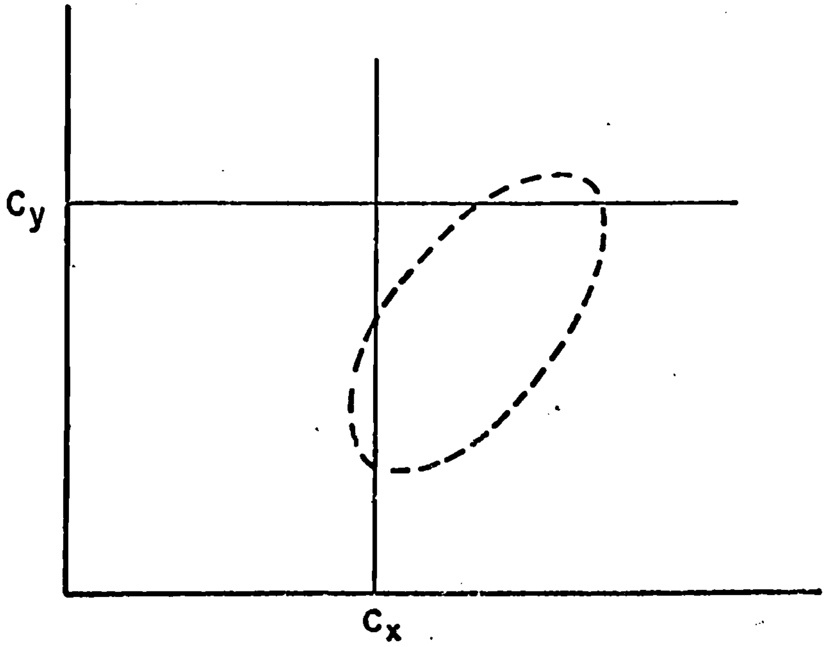

Figure 4. Norm-referenced correlation is positive; criterion-referenced correlation is negative.