DOCUMENT RESUME

ED 042 802 24 TM 000 070

AUTHOR	Livingston, Samuel A.
TITLE	The Reliability of Criterion-Referenced Measures.
INSTITUTION	Johns Hopkins Univ., Baltimore, Md. Center for the Study of Social Organization of Schools.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-6-1610
PUB DATE	Jul 70
GRANT	OEG-2-7-061610-0207
NOTE	20p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.10
DESCRIPTORS	Correlation, *Criterion Referenced Tests, Measurement Instruments, Norm Referenced Tests, Raw Scores, *Reliability, *Statistical Analysis, Statistics, *Test Reliability, *Tests, True Scores

ABSTRACT

The assumptions of the classical test-theory model are used to develop a theory of reliability for criterion-referenced measures-which parallels that for norm-referenced measures. It is shown that the Spearman-Brown formula holds for criterion-referenced measures and that the criterion-referenced reliability coefficient can be used to correct criterion-referenced correlations for attenuation. A formula is developed which expresses the criterion-referenced reliability coefficient in terms of the mean, variance, and norm-referenced reliability coefficient. The implications of the resulting formula are discussed. (Author/DG)

THE JOHNS HOPKINS UNIVERSITY

REPORT No. 73

THE RELIABILITY OF CRITERION-REFERENCED MEASURES

SAMUEL A. LIVINGSTON

JULY 1970

STAFF

John L. Holland, Director

James M. McPartland, Assistant Director

Virginia Bailey

Thelma Baldwin

Zahava D. Blum

Judith P. Clark

Karen C. Cohen

James S. Coleman

Robert L. Crain

David DeVries

Keith Edwards

Doris R. Entwisle

James Fennessey

Catherine J. Garvey

Ellen Greenberger

John T. Guthrie

Rubie Harris

Edward J. Harsch

Robert T. Hogan

Marian Hoover

Thomas Houston

Michael Inbar

Nancy L. Karweit

Judith Kennedy

Steven Kidder

Hao-Mei Kuo

Samuel Livingston

Edward L. McDill

Rebecca J. Muraro

Jeanne O'Connor

Martha O. Roseman

Peter H. Rossi

Joan Sauer

Leslie Schnuelle

Christine Schulze

Aage B. Sørensen

Annemette Sørensen

Julian C. Stanley

Clarice S. Stoll

Mary Viernstein

Murray A. Webster

Barbara J. Williams

Phyllis K. Wilson

The Reliability of Criterion-Referenced Measures

Grant No. -- OEG-2-7-061610-0207

Samuel A. Livingston

July, 1970

Report No. 73

Published by the Center for the Study of Social Organization of Schools, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education, and no official endorsement by the Office of Education should be inferred.

The Johns Hopkins University

Baltimore, Maryland

Acknowledgement

I thank Julian C. Stanley for the advice and instruction which made this paper possible.

The Reliability of Criterion-Referenced Measures

Abstract

The assumptions of the classical test-theory model are used to develop a theory of reliability for criterion-referenced measures which parallels that for norm-referenced measures. Tho criterion-referenced reliability coefficient is expressed in terms of the mean, variance, and norm-referenced reliability coefficient, and the implications of the resulting formula are discussed.

The Reliability of Criterion-Referenced Measures

"Criterion-referenced" is a term first used by Glaser (1963) to refer to measures that "depend on an absolute standard of quality." Thus criterion-referenced measures differ from "norm-referenced" measures, which depend on a relative standard. Criterion-referenced (CR) measures compare the student's performance with a fixed standard, while norm-referenced (NR) measures compare his performance with the performance of a norm group.

Popham and Husek (1969) have written that "the typical indices of internal consistency are not appropriate for criterion-referenced tests," and, at first glance, this point would seem so obvious as to be irrefutable. Since reliability theory is based on the existence of differences among the true scores of examinees, and CR measures are intended to apply to situations in which there may be no such differences, the two concepts would seem to be incompatible. Yet, with a few appropriate modifications, the classical theory of test reliability can be applied to criterion-referenced measures in a way that closely parallels its traditional application to norm-referenced measures.

The basis for these modifications is a simple substitution. Consider the basic distinction between NR and CR measures. When we use NR measures, we are interested in the extent to which an individual score deviates from the mean score of a norm group. When we use CR measures, we are interested in the extent to which an individual score deviates from a fixed standard, the criterion. To adapt traditional norm-referenced reliability indices to CR measures, one need only substitute the criterion score for the mean score of the norm group and redefine the various indices accordingly.

Variance, Covariance, and Correlation

How can we redefine the variance of scores on a CR test? The variance of a set of scores is the mean squared deviation of the scores from the group mean. Since we are interested not in the deviation of scores from the mean but in their deviation from the criterion, we can use, in place of the variance, the mean squared deviation of the scores from the criterion:

$$(1)\quad D_x^2 = E_p(X_{pf} - C_x)^2$$

where D_x^2 denotes the mean squared deviation of the X-measures from C_x , X_{pf} is the obtained score of person p on form f , C_x is the criterion, and E_p indicates the expected value over persons.

Since the concepts of covariance and correlation depend on differences in scores, they, too, will have to be redefined. In place of covariance, we have a mean product of deviations:

$$(2)\quad D_{xy} = E_p(X_{pf} - C_x)(Y_{pf} - C_y)$$

The criterion-referenced correlation coefficient can then be defined as

$$(3) \quad \rho_c(X , Y) = \frac{D_{xy}}{D_x D_y} .$$

ρ_c is a product-moment correlation based on moments about the arbitrary origins C_x and C_y , rather than about the means. The Pearson product-moment correlation, which will be referred to in this paper as the norm-referenced correlation ρ_N , is thus a special case of ρ_c (with some special properties which do not generalize to other cases of ρ_c).

Definitions, Assumptions, and Basic Theorems

Since the criterion is chosen without reference to the distribution of scores, we can define the criterion of a sum of measurements in any way we choose. However, in order to construct indices of reliability which parallel those for norm-referenced measurement, we will have to define the criterion of a sum of measures as the sum of their criteria:

$C_{(X + Y)} = C_x + C_y$. More generally,

$$(4) \quad C_{\left(\sum_{i=1}^{n} X_i\right)} = \sum_{i=1}^{n} C_{x_i} .$$

It follows that

(5) $C_{(nX)} = nC_x$.

True scores and errors of measurement are defined exactly as for NR measures:

$$T_p = E_f(X_{pf}) \quad \text{and} \quad e_{pf} = X_{pf} - T_p \ .$$

That is, the true score of person p equals the expected value (over forms) of his obtained score; his error of measurement on a given form is the difference between his obtained score on that form and his true score.

The concept of true-score variance must be replaced by the mean squared deviation of true scores from the criterion:

(6) $D_t^2 = E_p(T_p - C_x)^2$.

Classical test theory assumes that errors of measurement on separate measures do not covary over persons or over forms; the same assumptions can be made for CR measures:

$$E_p(e_{pf}e_{p'f}) = 0 \ ; \quad E_f(e_{pf}e_{pf'}) = 0 \ .$$

Classical test theory also assumes that errors of measurement do not covary with true scores on the same or on other measures. It follows that errors of measurement do not covary with the deviation of true scores from the criterion:

$$(7)\quad E_p[e_{pf}(T_p - C_x)] = E_p(e_{pf}T_p - e_{pf}C_x)$$

$$= E_p(e_{pf}T_p) - C_xE_p(e_{pf}) = 0 - 0 = 0 .$$

We can now provc a theorem analogous to tha theorem for NR measures which states that the variance (over persons) of obtained scores equals the variance of true scores plus the variance of errors of measurement. For CR measures, the theorem states that the mean squared deviation of obtained scores from the criterion equals the mean squared deviation of true scores from the criterion, plus the variance of errors of measurement. The proof of this latter theorem i3 as follows:

$$(8)\quad D_x^2 = E_p(X_{pf} - C_x)^2 = E_p[(T_p + e_{pf}) - C_x]^2$$

$$= E_p[(T_p - C_x) + e_{pf}]^2$$

$$= E_p(T_p - C_x)^2 + E_p(e_{pf})^2 + 2E_p[e_{pf}(T_p - C_x)]$$

$$= D_t^2 + \sigma_e^2 + 2(0)$$

$$= D_t^2 + \sigma_e^2 .$$

The Reliability Coefficient

Lord and Novick (1968, p. 61) define the reliability coefficient for norm-referenced measures as the squared correlation between true scores and obtained scores. We can follow their example and define the criterion-referenced reliability coefficient as the squared CR correlation between true scores and obtained scores:

$$D_{tx} = E_p(T_p - C_x)(X_{pf} - C_x)$$

$$= E_p(T_p - C_x)[(T_p + e_{pf}) - C_x]$$

$$= E_p(T_p - C_x)[(T_p - C_x) + e_{pf}]$$

$$= E_p(T_p - C_x)^2 + E_p[e_{pf}(T_p - C_x)]$$

$$= D_t^2 + 0 = D_t^2 \ .$$

Therefore,

$$(9) \quad \rho_c^2(T \ , X) = \frac{(D_t^2)^2}{D_t^2 D_x^2} = \frac{D_t^2}{D_x^2} \ .$$

This result shows that the reliability coefficient of a criterion-referenced measure can be interpreted as a ratio of mean squared deviations from the criterion, just as the reliability coefficient of a norm-referenced measure can be interpreted as a ratio of variances.

We can define parallel measurements just as for NR measures, with the additional requirement that parallel measurements have equal criteria. Then two criterion-referenced measures X_1 and X_2 are parallel if and only if the following conditions hold:

$T_{p_1} = T_{p_2}$ for all p ;

$\sigma^2_{e_1} = \sigma^2_{e_2}$; and $C_1 = C_2$.

We can then show that the correlation of two parallel measures X and X' is equal to the reliability coefficient of X . The proof is as follows (the notation has been simplified to avoid two levels of subscripts):

$$\rho_c(X , X') = \frac{D_{xx'}}{D_x D_{x'}} .$$

Expanding the numerator,

$$D_{xx'} = E_p(X_p - C)(X'_p - C)$$

$$= E_p[(T_p + e_p) - C][(T_p + e'_p) - C]$$

$$= E_p[(T_p - C) + e_p][(T_p - C) + e'_p]$$

$$= E_p(T_p - C)^2 + E_p(e_p e'_p) + E_p[e_p(T_p - C)] + E_p[e'_p(T_p - C)]$$

$$= D^2_t + 0 + 0 + 0 = D^2_t .$$

From equation (8), $D^2_x = D^2_t + \sigma^2_e$;

therefore,

$$\rho_c(X\ ,\ X') = \frac{D_t^2}{\sqrt{(D_t^2 + \sigma_e^2)(D_t^2 + \sigma_{e'}^2)}}\ .$$

But, by the definition of parallel measurements stated earlier, $\sigma_e^2 = \sigma_{e'}^2$. Therefore,

$$(10)\quad \rho_c(X\ ,\ X') = \frac{D_t^2}{\sqrt{(D_t^2 + \sigma_e^2)^2}} = \frac{D_t^2}{D_t^2 + \sigma_e^2} = \frac{D_t^2}{D_x^2} = \rho_c^2(T\ ,\ X)\ .$$

The Spearman-Brown Formula

Does the Spearman-Brown formula hold for criterion-referenced measures? It does, and its derivation for CR measures parallels that for NR measures. Suppose we want to know the criterion-referenced reliability of a sum of n parallel measurements. By the definition of parallel measurements, all n criteria are equal; therefore, from equation (4), the criterion for the sum is nC_x .

The mean squared deviation of the true scores is

$$(11)\quad D^2_{(\Sigma T)} = D^2_{(nT)} = E_p(nT_p - nC_x)^2 = E_p[n(T_p - C_x)]^2$$

$$= n^2E_p(T_p - C_x)^2 = n^2D_t^2\ .$$

The mean squared deviation of the obtained scores is

$$D^2_{(\Sigma X)} = E_p(\sum_f^n X_{pf} - nC_x)^2$$

$$= E_p(nT_p + \sum_f^n e_{pf} - nC_x)^2$$

$$= E_p[n(T_p - C_x) + \sum_f^n e_{pf}]^2$$

$$= E_p[n^2(T_p - C_x)^2] + E_p(\sum_f^n e_{pf})^2 + E_p[2n(T_p - C_x)\sum_f^n e_{pf}]$$

$$= n^2E_p(T_p - C_x)^2 + E_p(\sum_f^n e_{pf}^2 + \sum_{f\neq f'}^n\sum^n e_{pf}e_{pf'}) + 2nE_p[(T_p - C_x)\sum_f^n e_{pf}] \ .$$

The first term equals $n^2D_t^2$. The second term equals

$$E_p(\sum_f^n e_{pf}^2) + E_p(\sum_{f\neq f'}^n\sum^n e_{pf}e_{pf'}) = \sum_f^n E_p(e_{pf}^2) + \sum_{f\neq f'}^n\sum^n E_p(e_{pf}e_{pf'}) = \sum_f^n \sigma_{e_f}^2 + 0 = n\sigma_e^2 \ .$$

The third term equals

$$2nEp[\sum_f^n e_{pf}(T_p - C_x)] = 2n\sum_f^n E_p[e_{pf}(T_p - C_x)] = 0 \text{ , by equation (7).}$$

Therefore,

$$(12) \quad D_{(\Sigma X)}^2 = n^2D_t^2 + n\sigma_e^2 \ .$$

Then the CR reliability coefficient of the sum, by equation (9), equals

$$(13) \quad \frac{D_{(\Sigma T)}^2}{D_{(\Sigma X)}^2} = \frac{n^2D_t^2}{n^2D_t^2 + n\sigma_e^2} = \frac{nD_t^2}{nD_t^2 + \sigma_e^2} = \frac{nD_t^2}{(n-1)D_t^2 + (D_t^2 + \sigma_e^2)}$$

$$\frac{nD_t^2}{(n-1)D_t^2 + D_x^2} = \frac{n\left(\dfrac{D_t^2}{D_x^2}\right)}{(n-1)\left(\dfrac{D_t^2}{D_x^2}\right) + 1} = \frac{n\rho_c^2(T \ , \ X)}{1 + (n-1)\rho_c^2(T \ , \ X)} \ .$$

Correction for Attenuation

The CR reliability coefficient can be used to correct CR correlations for attenuation. Again, the formula and its derivation parallel those for NR measures. First we must prove that $D_{T_x T_y} = D_{xy}$:

$$
\begin{aligned}
(14) \quad D_{xy} &= E_p(X_p - C_x)(Y_p - C_y) \\
&= E_p(T_x + e_x - C_x)(T_y + e_y - C_y) \\
&= E_p[(T_x - C_x) + e_x][(T_y - C_y) + e_y] \\
&= E_p(T_x - C_x)(T_y - C_y) + E_p[e_x(T_y - C_y)] + E_p[e_y(T_x - C_x)] + E_p(e_x e_y) \\
&= D_{T_x T_y} + 0 + 0 + 0 \\
&= D_{T_x T_y} \ .
\end{aligned}
$$

By the definition of CR correlation, equation (3),

$$\rho_c(T_x \, , \, T_y) = \frac{D_{T_x T_y}}{D_{T_x} D_{T_y}} \ .$$

But $D_{T_x T_y} = D_{xy}$, by equation (14).

And, since $\rho_c^2(T_x \, , \, X) = \dfrac{D^2_{T_x}}{D^2_x}$, then

$D^2_{T_x} = D^2_x \cdot \rho^2_c(T_x , X)$, and

$D_{T_x} = D_x \cdot \rho_c(T_x , X)$.

Similarly, $D_{T_y} = D_y \cdot \rho_c(T_y , Y)$.

Then

$$(15) \quad \rho_c(T_x , T_y) = \frac{D_{xy}}{D_x \cdot \rho_c(T_x , X) \cdot D_y \cdot \rho_c(T_y , Y)}$$

$$= \frac{1}{\rho_c(T_x , X) \cdot \rho_c(T_y , Y)} \cdot \frac{D_{xy}}{D_x D_y}$$

$$= \frac{\rho_c(X , Y)}{\rho_c(T_x , X) \cdot \rho_c(T_y , Y)} .$$

Computing Criterion-Referenced Indices from Norm-Referenced Indices

Suppose we have computed (or have a computer program for computing) the traditional norm-referenced indices for a set of scores: the mean, variance, and estimated reliability coefficient. Can we use these norm-referenced indices to compute criterion-referenced indices, including the reliability coefficient, without having to refer back to each student's response to each item? The answer is yes; in fact, we can compute criterion-referenced indices for this set of scores with any criterion we choose to specify.

Let the mean, variance, and norm-referenced reliability coefficient be represented by μ_x , σ^2_x , and $\rho^2_N(T , X)$. Then the mean squared deviation of obtained scores from the criterion can be expressed as follows:

$$
\begin{aligned}
(16)\quad D_x^2 &= E_p(X_{pf} - C_x)^2 = E_p[(X_{pf} - \mu_x) + (\mu_x - C_x)]^2 \\
&= E_p(X_{pf} - \mu_x)^2 + E_p(\mu_x - C_x)^2 + 2E_p(X_{pf} - \mu_x)(\mu_x - C_x) \\
&= \sigma_x^2 + (\mu_x - C_x)^2 + 2(\mu_x - C_x)E_p(X_{pf} - \mu_x) \\
&= \sigma_x^2 + (\mu_x - C_x)^2 \ .
\end{aligned}
$$

A similar derivation holds for the mean squared deviation of true scores from the criterion. The result is

$$
(17)\quad D_t^2 = \sigma_t^2 + (\mu_t - C_x)^2 = \rho_N^2(T\ ,\ X)\sigma_x^2 + (\mu_x - C_x)^2 \ .
$$

The mean product of deviations for two CR measures can be expressed in terms of the means, criteria, and covariance of the two measures:

$$
\begin{aligned}
(18)\quad D_{xy} &= E_p(X_{pf} - C_x)(Y_{pf} - C_y) \\
&= E_p[(X_{pf} - \mu_x) + (\mu_x - C_x)][(Y_{pf} - \mu_y) + (\mu_y - C_y)] \\
&= E_p(X_{pf} - \mu_x)(Y_{pf} - \mu_y) + E_p(\mu_x - C_x)(\mu_y - C_y) \\
&\qquad + E_p(\mu_x - C_x)(Y_{pf} - \mu_y) + E_p(\mu_y - C_y)(X_{pf} - \mu_x) \\
&= \sigma_{xy} + (\mu_x - C_x)(\mu_y - C_y) + (\mu_x - C_x)E_p(Y_{pf} - \mu_y) \\
&\qquad + (\mu_y - C_y)E_p(X_{pf} - \mu_x) \\
&= \sigma_{xy} + (\mu_x - C_x)(\mu_y - C_y) \ .
\end{aligned}
$$

Then the criterion-referenced correlation coefficient can be expressed in terms of norm-referenced indices:

$$(19)\quad \rho_c(X , Y) = \frac{D_{xy}}{D_x D_y} = \frac{\rho_N(X , Y)\sigma_x\sigma_y + (\mu_x - C_x)(\mu_y - C_y)}{\sqrt{[\sigma_x^2 + (\mu_x - C_x)^2][\sigma_y^2 + (\mu_y - C_y)^2]}}$$

Since we can express the mean squared deviation of obtained scores and that of true scores in terms of norm-referenced indices, we can do the same for their ratio, which is the criterion-referenced reliability coefficient:

$$(20)\quad \rho_c^2(T , X) = \frac{D_t^2}{D_x^2} = \frac{\rho_N^2(T , X)\sigma_x^2 + (\mu_x - C_x)^2}{\sigma_x^2 + (\mu_x - C_x)^2} .$$

Implications of Criterion-Referenced Reliability

Consider the implications of equation (20). As the NR reliability coefficient increases, the CR reliability coefficient increases. When the NR reliability coefficient equals 1.00, the CR reliability coefficient also equals 1.00. In fact, the CR reliability coefficient is always at least as large as the NR reliability coefficient. The two reliability coefficients will be equal whenever the mean score falls exactly at the criterion.

The further from the criterion the mean score falls, the greater the CR reliability coefficient. The reason for this relationship is that the mean of the obtained scores is equal to the mean of the true scores--the point from which the sum of squared deviations of the individual true scores is the smallest it can be. The farther from this point the criterion lies, the more reliable information one has about the deviation of all the individual true scores from the criterion. For this reason, NR reliability can be considered a special case of CR reliability--the case in which the mean and the criterion are equal and the reliability of the test is minimized.

Another way to think about the relationship between the mean, the criterion, and the CR reliability of the test is in terms of D_t^2 and σ_e^2 . From equations (8), (9), and (17),

$$(21) \quad \rho_c^2(T\ ,\ X) = \frac{D_t^2}{D_t^2 + \sigma_e^2} = \frac{\sigma_t^2 + (\mu_x - C_x)}{\sigma_t^2 + (\mu_x - C_x) + \sigma_e^2} .$$

Increasing the distance between the mean and the criterion increases the mean squared deviation of the true scores from the criterion, without any increase in the error variance. As a result, the CR reliability increases.

How is CR reliability affected by a decrease in the variance of obtained scores? The answer depends on the nature of the decrease in variance. If the NR reliability remains constant--that is, if true-score variance and error variance decrease in the same proportion--the CR reliability will increase. The effect is the same as that of increasing $(\mu_x - C_x)^2$ while holding σ_t^2 and σ_e^2 constant.

However, a decrease in obtained-score variance is usually accompanied by a decrease in the NR reliability coefficient. What usually happens is that the true-score variance decreases while the error variance remains constant. In this case, of course, CR reliability will decrease.

What about the case of a mastery test, on which all the students are expected to get perfect scores? If they all get perfect scores, does the test have no reliability? No, because the criterion is a point selected to divide the scores above it from those below. Therefore, the criterion for a mastery test is not a perfect score; it is a perfect score minus some small fraction of an item. If all the students get perfect scores, the variances in formula (21) will equal zero. Since there will still be the difference of a fraction of an item between the mean and the criterion, the CR reliability will equal 1.00.

There is one theoretically possible case for which CR reliability is undefined: that in which all the students obtain scores exactly at the criterion level. In this case both numerator and denominator in any of the formulas for the CR reliability coefficient would equal zero. But this case is not a practical possibility; if the lowest passing score is $\underline{k}$ items, the criterion is actually $\underline{k}$ minus some fraction of an item. However, it is possible for a test to have CR reliability equal to zero. This will happen when the mean score falls exactly at the criterion and the NR reliability equals zero.

References

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

Lord, Frederic M. and Novick, Melvin R. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley, 1968.

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