CRITERION-REFERENCED RELIABILITY: A COMPARISON OF FIVE METHODS OF ESTIMATING THE LIVINGSTON COEFFICIENT

Hubert T. Lovett, Mississippi State University

1. INTRODUCTION

The reliability of a mental measurement can be viewed as a measure of the degree to which the measurement discriminates between individual performance and some point $\underline{C}$ on the score scale. In the case of a norm-referenced test, the point $\underline{C}$ is set equal to the population mean for the test. In the case of a criterion-referenced test, on the other hand, the point $\underline{C}$ is determined without regard to group performance; it is generally a minimum level of acceptable performance, or cutting score.

Estimation of the reliability of norm-referenced tests has been well established in theory, and the theory has generated methodology sufficient for most educational situations (8). However, criterion-referenced tests have become popular only in the last few years, and only recently have psychometricians attempted to develop theoretical explanations of the reliability of criterion-referenced tests (2, 6, 11). The purpose of this study was to compare the relative validity of five methods of estimating the reliability, as defined by Livingston (6), of criterion-referenced tests.

2. THE LIVINGSTON COEFFICIENT

Livingston's explanation of the reliability of criterion-referenced tests depends upon defining observed and true variance as the expected squared deviation of the respective score from the criterion, $\underline{C}$, rather than from the population mean (6, 9). Observed variance so defined can be partitioned thus,

$$\underline{E}(\underline{D}^2_{x_i}) = \underline{E}(\underline{D}^2_{T_i}) + \sigma^2_{e_i}, \qquad (2.1)$$

where $\underline{E}$ indicates the expected value, $\underline{D}_{x_i}$, the deviation of the observed score of the ith person from $\underline{C}$, $\underline{D}_{T_i}$, the deviation of the true score of the ith person from $\underline{C}$, and $\sigma^2_{e_i}$, the expected squared deviation of the ith observed score from the corresponding true score. Criterion-referenced reliability is then defined by analogy to norm-referenced reliability as $\underline{R}_{cc}$

$$= \underline{E}(\underline{D}^2_{T_i})/\underline{E}(\underline{D}^2_{x_i}) = E(\underline{D}^2_{T_i})/(\underline{E}(\underline{D}^2_{T_i}) + \sigma^2_{e_i}) \qquad (2.2)$$

where $\underline{R}_{cc}$ is the criterion-referenced reliability.

3. ESTIMATING $\underline{R}_{cc}$

Three estimates of $\underline{R}_{cc}$ were taken from existing literature, and two new methods were developed. The three techniques taken from existing literature require that all the test items be dichotomously (zero or one) scored. The two new methods, based on the analysis of variance, require only that all items be parallel measurements.

3.1 The Binomial Method

Lovett (9) derived a formula based on the binomial distribution which gives an estimate of $\underline{R}_{cc}$ given that, for each person taking the test, the probability of correctly answering a question is constant over all questions:

$$\underline{r}_{cc}(\text{Binomial}) = 1 - (\underline{k}\Sigma\underline{X}_i - \Sigma\underline{X}^2_i)/((\underline{k} - 1)\Sigma(\underline{X}_i - \underline{C})^2), \qquad (3.1.1)$$

where $\underline{r}_{cc}$ is a sample estimate of $\underline{R}_{cc}$, $\underline{X}_i$, the observed score for the ith person, $\underline{k}$, the number of questions on the test, and $\underline{C}$, the criterion.

3.2 Analysis of Variance

The analysis of variance has been used to estimate the reliability of norm-referenced tests (Hoyt, 1941; Winer, 1971). In estimating the reliability of a test with the analysis of variance, the testing situation is conceptualized as an $\underline{n}$-persons-x-$\underline{k}$-items design, the observation for cell$_{ij}$ is the score for the ith person on the jth test item. It is assumed that there is no item-x-person interaction. In extending the model to the criterion-referenced situation the grand mean is partially replaced in the score model by $\underline{C}/\underline{k}$, thus,

$$\underline{X}_{ij} = (\underline{C}/\underline{k}) + ((\underline{X}_i/\underline{k}) - (\underline{C}/\underline{k})) + ((\underline{X}_j/n) - \overline{\underline{X}}) + e_{ij}, \qquad (3.2.1)$$

where $\underline{X}_{ij}$ is the score for the ith person on the jth item, $\underline{X}_j$, the sum for item j, $\overline{\underline{X}}$, the grand mean, and $\underline{e}_{ij}$, an error of measurement. The expected values for the mean squares, person and error, can be shown to be

$$\underline{E}(\underline{MS}_p) = \sigma^2_{e_i} + \underline{E}(\underline{D}^2_{T_i}), \qquad (3.2.2)$$

and

$$E(\underline{MS}_e) = \sigma^2_{e_i}, \qquad (3.2.3)$$

where $\underline{MS}_p$ is the mean square person and $\underline{MS}_e$, the error, person-x-item, mean square. From (2.1), (2.2), (3.2.2), and (3.2.3) it follows that

$$\underline{R}_{cc} = (\underline{E}(\underline{MS}_p) - \underline{E}(\underline{MS}_e))/\underline{E}(\underline{MS}_p). \quad (3.2.4)$$

Therefore, an estimate of $\underline{R}_{cc}$ is given by

$$\underline{r}_{cc}(\text{ANOVA}) = (\underline{MS}_p - \underline{MS}_e)/\underline{MS}_p \quad (3.2.5)$$

which is equivalent to

$$\underline{r}_{cc}(\text{ANOVA}) = (\underline{F} - 1)/\underline{F}, \quad (3.2.6)$$

where $\underline{F} = \underline{MS}_p/\underline{MS}_e$. $\underline{MS}_e$ is defined by the formula

$$\underline{MS}_e = (\Sigma\Sigma(\underline{X}_{ij} - (\underline{C}/\underline{k}))^2 - \underline{n}\Sigma((\underline{X}_j/\underline{n}) - \overline{\underline{X}})^2 - \underline{k}\Sigma((\underline{X}_i/\underline{k}) - (\underline{C}/\underline{k}))^2)/((\underline{k} - 1)(\underline{n} - 1)), \quad (3.2.7)$$

and $\underline{MS}_p$ is defined by the formula

$$\underline{MS}_p = (\underline{k}/\underline{n})\Sigma((\underline{X}_i/\underline{k}) - (\underline{C}/\underline{k}))^2. \quad (3.2.8)$$

It should be noted that the degrees of freedom for $\underline{MS}_p$ is $\underline{n}$ instead of the usual $\underline{n} - 1$, because $\overline{\underline{X}}$ has been replaced by $\underline{C}/\underline{k}$ which is independent of the scores and, therefore, does not represent a constraint on the value of $\underline{MS}_p$.

3.3 Corrected Analysis of Variance

Winer (1971) pointed out that the reliability in formula (3.2.6) will be biased. To obtain an unbiased estimate of $\underline{R}_{cc}$ the following correction is necessary:

$$\underline{F}' = (\underline{MS}_p)/(\underline{m}(\underline{MS}_e)), \quad (3.3.1)$$

where $\underline{m} = \underline{n}(\underline{k} - 1)/(\underline{n}(\underline{k} - 1) - 2)$. The formula for $\underline{r}_{cc}$ is then

$$\underline{r}_{cc}(\text{ANOVA Corrected}) = (\underline{F}' - 1)/\underline{F}' \quad (3.3.2)$$

3.4 Kuder and Richardson's Formulae 20 and 21

Livingston (6) and Mehrens and Lehmann (10) suggested that the reliability of a criterion-referenced test be estimated by finding the reliability of the test as though it were a norm-referenced test and adjusting the reliability to the criterion-referenced situation thus:

$$\underline{r}_{cc} = (\underline{r}_{xx}\underline{S}_x^2 + (\overline{\underline{X}} - \underline{C})^2)/(\underline{S}_x^2 + (\overline{\underline{X}} - \underline{C})^2) \quad (3.4.1)$$

where $\underline{r}_{xx}$ is any norm-referenced estimate of the reliability of the test, $\underline{S}_x^2$ is an estimate of the observed test variance around the population mean, and $\overline{\underline{X}}$ is an estimate of the test mean. Because of their popularity Kuder and Richardson's (5) formulae 20 and 21 were selected for use in this study.

Kuder and Richardson's formula 20 is as follows:

$$\underline{r}_{KR-20} = (\underline{k}/(k - 1)(1 - (\Sigma\underline{p}_j\underline{q}_j)/\underline{S}_x^2), \quad (3.4.2)$$

where $\underline{r}_{KR-20}$ is an estimate of the norm-referenced reliability of the test, $\underline{p}_j$ is the mean of the jth question and $\underline{q}_j = 1 - \underline{p}_j$. It is assumed that all interitem correlations are equal, and the matrix of interitem correlations has a rank of one. Substituting in (3.4.1) from (3.4.2) gives

$$\underline{r}_{cc}(\text{KR-20}) = (\underline{r}_{KR-20}\underline{S}_x^2 + (\overline{\underline{X}} - \underline{C})^2)/(\underline{S}_x^2 + (\overline{\underline{X}} - \underline{C})^2) \quad (3.4.3)$$

Kuder and Richardson's formula 21 requires the additional assumption that $\underline{p}_j$ is constant for all j. With this assumption $\underline{r}_{KR-20}$ can be reduced to $\underline{r}_{KR-21}$:

$$\underline{r}_{KR-21} = (\underline{k}/(\underline{k} - 1))(1 - (\overline{\underline{X}}(\underline{k} - \overline{\underline{X}})/kS_x^2)). \quad (3.4.4)$$

Substituting in (3.4.1) from (3.4.4) gives

$$\underline{r}_{cc}(\text{KR-21}) = (\underline{r}_{KR-21}\underline{S}_x^2 + (\overline{\underline{X}} - \underline{C})^2)/(\underline{S}_x^2 + (\overline{\underline{X}} - \underline{C})^2). \quad (3.4.5)$$

4. METHOD

4.1 Procedure

The following parameters were varied to form 1024 different cases: The number ($\underline{n}$) of persons taking the test was varied from 25 to 100 by increments of 25; the number of test items ($\underline{k}$) was varied from 20 to 80 by increments of 20; the criterion ($\underline{C}$) was varied from $(.6)\underline{k}$ to $(.9)\underline{k}$ by increments of $(.1)\underline{k}$; the population mean (μ) was varied from approximately ($\underline{C} - .09\underline{k}$) to approximately ($\underline{C} + .09\underline{k}$) by increments of approximately $.06\underline{k}$; the variance of true scores around μ (σ_T^2) was varied from approximately 9.00 to approximately 56.25 by incrementing σ_T by approximately 1.5.

The use of approximate limits and increments for μ and σ_T^2 resulted from the manner in which the true scores were formed. For each of the 1024 cases the pseudo-random number generator "Randn" (12) was used to form a set of n random, normal, true scores. The program allows exact specification of the mean and standard deviation. The program, however, does not allow for the specification of limits; therefore, some of the scores were outside the test limits: either larger than k or smaller than zero. Those larger than k were set equal to $k - b$, where b was a pseudo-random number, between zero and one, generated by the generator "Randu" (13). True scores smaller than zero were set equal to b. After bringing all true scores within the limits of the test, μ and σ_T^2 were re-calculated. It was deemed more important that the data conform to realistic test situations than that the increments and limits of μ and σ_T^2 be exact.

After the n true scores (T) were formed in each case, five n-persons-by-k-items, item-pattern matrixes were formed. The score for the ith person on the jth item in the kth matrix was one (indicating a correct answer) if $b_{ijk} < (T_i/k)$ and zero (indicating an incorrect answer) otherwise, where b was an array of pseudo-random numbers having a uniform distribution on the interval zero to one, formed by the generator "Randu" (13). The array b and the true scores were formed independently for each of the 1024 cases.

Because the true scores were known in each case, $E(D_{T_i}^2)$ could be calculated, and because of the method of forming the item-pattern matrixes, p_{ij}, the probability that the ith person would answer the jth question correctly, equaled T_i/k. Because the Ts and ps were known $\sigma_{e_i}^2$ could be found by a formula derived by Lord (7). R_{cc} could then be calculated for each case. Also, the method of forming the item-pattern matrixes assured that all of the assumptions of the five methods of estimating R_{cc} were met.

In each of the 1024 cases r_{cc}(Binomial) was calculated for the first item-pattern matrix; r_{cc}(ANOVA), for the second; and so on.

4.2 Analysis

For each of the five methods in all 1024 cases an error term was calculated, defined as

$$w = r_{cc} - R_{cc}, \qquad (4.2.1)$$

where w is the error in estimating R_{cc}. In each case the method having the smallest absolute w was given a rank of one; the one with the next smallest, a rank of two, and so on up to five. The ranks were then summed for each method over all 1024 cases. Using the summed ranks Friedman's Two-Way ANOVA by Ranks was used to test the hypothesis that the sum of ranks were constant across all five methods (3). There are no distributional assumptions associated with the Friedman test. A distribution-free, multiple-comparison procedure was used to test all pairwise contrast (3).

5. RESULTS AND DISCUSSION

Table one summarizes the results of the analysis. The chi square calculated as the test statistic of the Friedman test was 3071.24(df = 4, $P < .01$). The multiple comparison procedure revealed that an absolute difference of 232.90 between any two sum of ranks was significant at the .01 level. Therefore, the hypothesis that the five methods are equally valid estimates of R_{cc} was rejected. The multiple comparison procedure revealed no significant differences among r_{cc}(Binomial), r_{cc}(KR-20), and r_{cc}(KR-21). However these three methods did differ significantly from the two ANOVA methods, but r_{cc}(ANOVA) and r_{cc}(ANOVA Corrected) did not differ

1. Result of Analysis

	Method				
	Binomial	ANOVA	ANOVA Corrected	KR-20	KR-21
Mean w	-.004	-.866	-.859	-.005	-.003
Variance of w	.002	.613	.627	.001	.002
Conservative estimates	527	1024	1024	364	479
Non-interpretable Cases	1	391	383	0	0
Sum of Ranks	2009	4625	4590	2079	2057
Multiple Comparison*	A	B	B	A	A

*Two methods having the same letter are not significantly different at the .01 level. Two methods not having the same letter are significantly different at the .01 level.

significantly from each other.

Given approximately equal validity in two methods, there are some very strong arguments for preferring the method which will most likely yield a conservative estimate of $\underline{R}_{cc}$. Thus a count of conservative estimates was made, where a conservative estimate was defined as a case where $\underline{w}<0.0$. It was found that the proportion of conservative estimates was significantly ($\underline{p}<.01$) larger for $\underline{r}_{cc}$(Binomial) than for either $\underline{r}_{cc}$(KR-20) or $\underline{r}_{cc}$(KR-21). The standard, normal deviate $\underline{z}$ was the test statistic (1). In comparing $\underline{r}_{cc}$(Binomial) with $\underline{r}_{cc}$(KR-20), $\underline{z} = 7.27$. In comparing $\underline{r}_{cc}$(Binomial) with $\underline{r}_{cc}$(KR-21), $\underline{z} = 2.12$. When $\underline{r}_{cc}$(KR-20) was compared with $\underline{r}_{cc}$(KR-21), $\underline{z} = 5.16$.

It is also desirable to avoid methods which are likely to yield non-interpretable results. A non-interpretable result was defined as a negative value for $\underline{r}_{cc}$. Table 1 shows that the frequency of non-interpretable results for the two ANOVA methods was very large, whereas $\underline{r}_{cc}$(Binomial), $\underline{r}_{cc}$(KR-20), and $\underline{r}_{cc}$(KR-21) had only one among them. The item-pattern matrix for which $\underline{r}_{cc}$(Binomial) yielded a non-interpretable result was fed into the $\underline{r}_{cc}$(KR-20) and $\underline{r}_{cc}$(KR-21) subroutines, and both of them also yielded non-interpretable results. This indicates that any of the methods can yield non-interpretable results, but with $\underline{r}_{cc}$(Binomial), $\underline{r}_{cc}$(KR-20), and $\underline{r}_{cc}$(KR-21), the situations in which non-interpretable results will occur are very rare.

6. CONCLUSIONS

The results tend to support the following conclusions: The five methods are not equally valid estimates of $\underline{R}_{cc}$--$\underline{r}_{cc}$(Binomial), $\underline{r}_{cc}$(KR-20), and $\underline{r}_{cc}$(KR-21) being the more valid estimates of $\underline{R}_{cc}$. The "valid" methods are not equally conservative, $\underline{r}_{cc}$(Binomial) being the most conservative of the three. The data did not permit conclusions about cases where the assumptions of the various methods are not met, or cases where test items are not dichotomously scored (zero and one). Finally, no attempt was made to identify relationships between the parameters, which were varied to form the 1024 cases, and the relative validity of the five methods.

REFERENCES

(1) Glass, G. V. and Stanley, J. C., Statistical Methods in Education and Psychology, Englewood Cliffs: Prentice-Hall, 1970.

(2) Hambleton, R. K. and Novick, M. R., "Toward an Integration of Theory and Method for Criterion-referenced Tests," Journal of Educational Measurement, 10(Fall 1973), 159-170.

(3) Hollander, M. and Wolfe, D. A., Nonparametric Statistical Methods, New York: John Wiley, 1973.

(4) Hoyt, C. J., "Test Reliability Estimated by Analysis of Variance," Psychometrika, 6(June 1941), 153-160.

(5) Kuder, G. F. and Richardson, M. W., "The Theory of the Estimation of Test Reliability," Psychometrika, 2(June 1937), 151-160.

(6) Livingston, S. A., "Criterion-referenced Applications of Classical Test Theory," Journal of Educational Measurement, 9(Spring 1972), 13-26.

(7) Lord, F. M., "Do Tests of the Same Length Have the Same Standard Error of Measurement," Educational and Psychological Measurement, 17(November 1957), 510-521.

(8) Lord, F. M. and Novick, M. R., Statistical Theories of Mental Test Scores, Reading, Mass.: Addison-Wesley, 1968.

(9) Lovett, H. T., "Elaboration and Application of a Theory of Criterion-referenced Reliability," Paper read at the meeting of the Southeastern Psychological Association, Atlanta, 1975.

(10) Mehrens, W. A. and Lehman, I. J., Measurement and Evaluation in Education and Psychology, New York: Holt, Rinehart and Winston, 1973.

(11) Popham, W. J. An Evaluation Guidebook: A Set of Practical Guidelines for the Educational Evaluator, Los Angeles: The Instructional Objectives Exchange, 1972.

(12) Randn, in Large-scale Systems Math-pack Programmers Reference, New York: Univac Division of Sperry Rand, 1973, Section 14, 8-12.

(13) Randu, in Large-scale Systems Math-pack Programmers Reference, New York: Univac Division of Sperry Rand, 1973, Section 14, 4-7.

(14) Winer, B. J., Statistical Principles in Experimental Design, New York: McGraw-Hill, 1971.