DOCUMENT RESUME

ED 120 229 TM 005 188

AUTHOR Subkoviak, Michael J.
TITLE Estimating Reliability from a Single Administration of a Mastery Test.
PUB DATE Apr 76
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Criterion Referenced Tests; *Mathematical Models; Scores; Statistical Analysis; *Test Reliability
IDENTIFIERS *Mastery Testing

ABSTRACT

A number of different definitions and indices of reliability for mastery tests have recently been proposed in an attempt to cope with possible lack of score variability that attenuates traditional coefficients. One promising index that has been suggested is the proportion of students in a group that are consistently assigned to the same mastery state across two testings. The present paper proposes a single test administration method of obtaining such an estimate. (Author)

Estimating Reliability From a Single Administration of a Mastery Test

Michael J. Subkoviak

The University of Wisconsin

Running Head: Single-Administration Reliability of a Mastery Test

Estimating Reliability From a Single Administration of a Mastery Test

A *mastery test* is one in which the range of possible scores is partitioned into k nonoverlapping intervals that define various levels of student mastery. The familiar pass-fail test with a criterion of 75 percent correct is an example of such a criterion-referenced (CR) test, where $k = 2$. Since mastery tests are often used in conjunction with instructional programs that maximize the number of students attaining the highest mastery states and minimize the variability of test scores, the classical correlation between scores on parallel tests (or equivalently, the ratio of true to observed variance) may be attenuated by lack of variability and thus is unsatisfactory as an indicator of CR reliability (Popham and Husek, 1969).

For this reason, Livingston (1972a,b,c, 1973) proposed the following index of CR reliability for the special case of $k = 2$ mastery states:

$$K^2(X,T) = \frac{\sigma^2(T) + (\mu-C)^2}{\sigma^2(X) + (\mu-C)^2} \qquad (1)$$

where X and T are observed and true scores respectively, μ is the mean score and C is the criterion score. In words, Equation 1 is the ratio of true variance plus $(\mu-C)^2$ to observed variance plus $(\mu-C)^2$. Thus, possible lack of score variability is compensated for by the addition of the squared distance between the mean and the criterion score. $K^2(X,T)$ increases as $(\mu-C)^2$ increases for fixed $\sigma^2(T)$ and $\sigma^2(X)$, which for certain distributions is indicative of the fact that assignment to mastery states is stable because scores do not cluster about C. However, for a symmetric, bimodal distribution $K^2(X,T)$ increases as C moves away from μ toward either of the two modes--even though assignment

to mastery states is more stable at μ than at the modes. Livingston's index has subsequently elicited criticism for a number of different reasons (Hambleton & Novick, 1973; Harris, 1972, 1973; Shavelson, Block & Ravelch, 1972; Raju, Note 1).

Harris (Note 2) thus proposed another coefficient for the case $k = 2$ -- the squared correlation between mastery state, scored 0 and 1, and total score. In analysis of variance terms, this is a strength of relationship index given by:

$$\mu_C^2 = \frac{SS_B}{SS_B + SS_W} \tag{2}$$

where SS_B and SS_W are between and within sums of squares from a one-way analysis of test score variance for the $k = 2$ groups defined by criterion C. However, as Harris notes, for symmetric distributions the maximum value of μ_C^2 occurs at $C = \mu$ when the proportions in the two groups equal one-half. In the case of a symmetric, unimodal distribution this implies that μ_C^2 is largest when $C = \mu$ is at the point of greatest score density and thus when assignment to mastery states is relatively unstable.

More recently, Hambleton and Novick (1973) have suggested that an index of CR reliability reflect the degree to which students are consistently assigned to the same mastery states across parallel test administrations, as measured by some coefficient of agreement across testings. Accordingly, Swaminathan, Hambleton and Algina (1974) proposed that the proportion of students consistently assigned to mastery states across two testings serve as an estimate, i.e.,

$$p_o = \sum_{i=1}^{k} p_{ii} \tag{3}$$

where p_{ii} is the proportion of students consistently assigned to the i^{th} mastery state across the two administrations. Actually, Swaminathan, Hambleton and Algina recommend that a simple function of p_o be used--namely, the proportion of consistent assignments beyond that expected by chance.

Most recently, Marshall and Haertel (Note 3) have suggested a single test administration coefficient of agreement estimate. Their method is one of computing the average p_o across all possible split-halves of a single test (denoted β because it is computationally analogous to the classical α coefficient) and then stepping up β by a Spearman-Brown type formula to obtain an estimate for the full-length test. Initial results based on simulated data seem to indicate that the Marshall-Haertel index behaves in a reasonable manner for different score distributions and criteria C (Marshall, Note 4), i.e., the coefficient increases and decreases appropriately as criterion C is variously set at points of light and heavy score concentration. However, the derivation of the index was basically empirical rather than theoretical, and thus many of its statistical properties are presently unknown. The purpose of the present paper is to propose an alternative, single-administration coefficient of agreement estimate that is based on well-known statistical theory.

The Mathematical Model

Let us begin by formally defining the *coefficient of agreement for an individual i* as the probability that i is assigned to the same mastery state on parallel tests X and X'. The model for the case of $k = 2$ mastery states defined by criterion score C is outlined here; but the model extends easily to $k > 2$ mastery states defined by multiple criteria $C_1, C_2, C_3 \ldots C_{k-1}$. Now, there are two ways that an individual i can be assigned to the same mastery state on parallel tests X and X' with criterion C: (1) $X_i \geq C$ and $X'_i \geq C$

indicating consistent mastery/mastery decisions and (2) $X_i < C$ and $X_i' < C$ indicating consistent nonmastery/nonmastery decisions. (There are also two ways that inconsistent decisions can arise: (1) $X_i \geq C$ and $X_i' < C$, and (2) $X_i < C$ and $X_i' \geq C$.) Thus the coefficient of agreement $p_C^{(i)}$ for person i can be written:

$$p_C^{(i)} = P(X_i \geq C, X_i' \geq C) + P(X_i < C, X_i < C) \tag{4}$$

where the terms on the right side of Equation 4 are the probability of consistent mastery/mastery and nonmastery/nonmastery decisions respectively. Equation 4 might be of interest to educators who want to determine the reliability of a mastery test for making decisions about a particular person in an individualized instructional program.

The coefficient of agreement for a group of N persons can now be defined as the mean of the individual coefficients:

$$p_C = \frac{\sum_{i=1}^{N} p_C^{(i)}}{N} = \frac{\sum_{i=1}^{N} [P(X_i \geq C, X_i' \geq C) + P(X_i < C, X_i' < C)]}{N} \tag{5}$$

Equation 5 is the sum of the probability of a consistent decision for each person i weighted by his or her probability of occurrence in the group, and so again represents the (group) probability of a consistent decision on parallel tests.

Let us now introduce two assumptions that make possible the estimation of the individual coefficient of Equation 4, and thus also the estimation of the group coefficient in Equation 5. The first assumption is that scores X_i and X_i' are independently distributed for a fixed person i (Lord and Novick,

1968). Under this assumption Equation 4 can be rewritten:

$$p_C^{(i)} = P(X_i \geq C) \cdot P(X_i' \geq C) + P(X_i < C) \cdot P(X_i' < C) \quad (6)$$

This assumption implies that the experience of taking test X does not affect the outcome on test X' for person i or vice versa; and its validity would depend upon the degree to which content and administration of the two tests are separate.

The second assumption is that the distributions of X_i and X_i' for a fixed person are identically binomial in form (Lord & Novick, 1968). This implies that each of the n items on a test is scored 0 and 1 and also that the experience of taking earlier test items does not affect outcomes on later items. Under this assumption Equation 6 simplifies to

$$\begin{aligned} p_C^{(i)} &= [P(X_i \geq C)]^2 + [P(X_i < C)]^2 \\ &= [P(X_i \geq C)]^2 + [1 - P(X_i \geq C)]^2 \end{aligned} \quad (7)$$

where

$$P(X_i \geq C) = \sum_{X_i=C}^{n} \binom{n}{X_i} p_i^{X_i} (1 - p_i)^{n-X_i} \quad (8)$$

The quantity p_i in Equation 8 is the true probability of a correct item response for person i, which can be estimated from his or her observed score X_i on a single test, e.g., $\hat{p}_i = X_i/n$. Thus, as illustrated in a later section, the probability of consistent classification for each person can be estimated by Equations 7 and 8 and for an entire group by Equation 5, using the data from a single test administration.

Furthermore, the marginal group probability of assignment to the mastery (nonmastery) state is the same for both X and X' under the assumption of identically distributed X_i and X_i'; and the group probability of a consistent de-

cision due to chance with criterion C is:

$$P_{chance/C} = P(X \geq C)\cdot P(X' \geq C) + P(X \leq C)\cdot P(X' \leq C)$$
$$= [P(X \geq C)]^2 + [1 - P(X \geq C)]^2 \quad (9)$$

where $$P(X \geq C) = \frac{\sum_{i=1}^{N} P(X_i \geq C)}{N} \quad (10)$$

Thus the group probability of a consistent decision beyond that expected by chance is given by the kappa coefficient (Cohen, 1960, 1968, 1972; Swaminathan, et al., 1974):

$$\kappa = \frac{P_C - P_{chance/C}}{1 - P_{chance/C}} \quad (11)$$

where P_C is given by Equation 5 and $P_{chance/C}$ is given by Equations 9 and 10.

At this point, it may be interesting to reflect on a more general mathematical model of which Equations 6, 7 and 8 constitute a special case. Figure 1

Insert Figure 1 about here

---

represents the outcomes over repeated, joint administrations of parallel tests X and X' to person i with criterion C. The hatched areas of Quadrants I and III represent consistent decisions. The essential problem is one of determining the proportion of the bivariate distribution that falls in these two quadrants, given data from a single test administration. The binomial model is a logical first choice because it is relatively simple and yet flexible enough to account for the change in different students' distributions of scores, as their true abilities vary from near the "floor" of a test through the midrange and to the "ceiling" (see Lord and Novick, 1968, p. 510). However, more complex models probably provide a more accurate description of reality in most testing

situations. For example, Equation 8 might be replaced by a compound binomial model (Lord & Novick, 1968, pp. 524-526):

$$P(X_i \geq C) = \sum_{X_i=C}^{n} \{\binom{n}{X_i} p_i^{X_i}(1-p_i)^{n-X_i} + A_i B(X_i)\} \quad (12)$$

A_i and $B(X_i)$ above are defined by:

$$A_i = \frac{n^2(n-1)S_\pi^2 p_i(1-p_i)}{2[M_X(n-M_X) - S_X^2 - nS_\pi^2]} \quad (13)$$

$$B(X_i) = \sum_{v=0}^{2} (-1)^{v+1}\binom{2}{v}\binom{n-2}{X_i-v} p_i^{X_i-v}(1-p_i)^{(n-2)-(X_i-v)} \quad (14)$$

In Equation 13 S_π^2 is the variance of the n item difficulties; M_X and S_X^2 are re-respectively the mean and variance of test scores for the group.

Estimating p_i

The computational process of the previous section is set in motion by estimating the probability of a correct item response p_i for each person from the observed data. $P(X_i \geq C)$ can then be computed by Equation 8 for the simple binomial model or Equation 12 for the compound binomial model, followed by Equations 7 and 5 or by Equations 9-11. The present section considers various ways of estimating p_i.

Simple Binomial Model

The traditional (maximum likelihood) estimator of p_i is the proportion of test items correctly answered by person i:

$$\hat{p}_i = X_i/n \quad (15)$$

where X_i is the number correct and n is the total number of items. Since the standard error of estimate in this case is $\sqrt{p_i(1-p_i)/n}$, Equation 15 should

lead to reasonably accurate results if $n > 40$, particularly if the mastery level of most students is well above (below) $p_i = .50$.

However, Equation 15 does not include certain collateral information, such as mean M_X and variance S_X^2, that is available in group testing situations. When the number of items n is small, the inclusion of such information is particularly important for obtaining better estimates of p_i than those given by Equation 15. For example, if the distribution of observed scores X for the group approximates some member of the negative hypergeometric family of unimodal distributions (see Lord & Novick, 1968, p. 519 for illustrations) a better estimate of p_i is given by the regression equation:

$$\hat{p}_i = \alpha_{21}\left(\frac{X_i}{n}\right) + (1-\alpha_{21})\left(\frac{M_X}{n}\right) \qquad (16)$$

where $\alpha_{21} = \frac{n}{n-1}\left[1 - \frac{M_X(n-M_X)}{nS_X^2}\right]$ is the Kuder-Richardson Formula 21 reliability coefficient (which is the squared correlation between observed and true score under the simple binomial model.)

Equation 16 assumes that person i is a member of a unimodal distribution with mean M_X and variance S_X^2. However, multimodal situations are possible if different grade levels are present or if the test items are designed to discriminate very sharply between masters and nonmasters. Blind use of Equation 16 in such situations can lead to erroneous p_i estimates because the means and variances of the separate populations may be very different from the mean and variance for the combined data. If the various populations are clearly distinguishable, a separate regression equation like (16) can be derived for each group. However, an estimation procedure for p_i that employs collateral information and yet is free of distributional assumptions has obvious advantages. One such estimate is given by (Lord and Novick, 1968, p. 514):

$$\hat{p}_X = 1 - \frac{n-X+1}{X} \frac{\phi(X-1)}{\phi(X)} \hat{p}_{X-1} \qquad (17)$$

where $\phi(X-1)$ and $\phi(X)$ are the relative frequency of $X-1$ and X in the combined group and $\hat{p}_{X-1}$ and $\hat{p}_X$ are the proportion estimates corresponding to scores of $X-1$ and X. Unfortunately, complexity is the price that one pays for the generality of Equation 17. Accurate estimation of $\phi(X-1)$ and $\phi(X)$ require a large sample of subjects. Additionally, since (17) represents n equations in $n+1$ unknowns, the researcher must specify one of the $\hat{p}_{X-1}$ values to set the estimation process in motion, e.g., if $X-1$ is a chance score on an m-option multiple choice test one might set $\hat{p}_{X-1} = 1/m$. See Lord (1959) for examples of the use of Equation 17. Further pursuit of simple, yet general, procedures for estimating p_i with small n is clearly indicated.

Compound Binomial Model

The procedures here are analogous to those above. If n is large the classical estimate of Equation 15 can be used.

However, the following regression estimate includes collateral information about the mean, variance and item difficulties for a unimodal distribution:

$$\hat{p}_i = \alpha_{20}\left(\frac{X_i}{n}\right) + (1-\alpha_{20})\left(\frac{M_X}{n}\right) \qquad (18)$$

where α_{20} is the Kuder-Richardson Formula 20 reliability coefficient (which is the squared correlation between observed and true scores under the compound binomial model).

Examples

In order to illustrate the computation of the individual and group coefficients $P_C^{(i)}$ and P_C, the simple binomial model will be applied first to a small set of stimulated data and then to real data.

As shown in Table 1, the true probability of a correct item response p_i

Insert Table 1 about here

was specified for each of $N = 10$ hypothetical subjects. An observed score X_i on an $n = 5$ item test was generated for each subject uaing p_i. For example, a random unit was drawn indicating Person 1's performance on each of the five items as follows: 9, 3, 6, 5, 2. Since Person 1'a probability of a correct response is $p_1 = .2$, Units 0-1 were scored as correct and Units 2-9 as incorrect, accordingly $X_1 = 0$ as shown in Table 1.

These single-administration X_i scores are used to estimate $P_C^{(i)}$ and P_C where $C = 4$. First the probability of a correct response $\hat{p}_i$ for each student is estimated by Equation 16. Next, $\hat{p}_i$ is substituted into Equation 8 with $C = 4$ and $n = 5$ to obtain $\hat{P}(X_i \geq 4)$ and its complement $1 - \hat{P}(X_i \geq 4)$ for each atudent. $\hat{P}(X_i \geq 4)$ and $1 - \hat{P}(X_i \geq 4)$ are squared and summed according to Equation 7 to provide an estimate $\hat{p}_4^{(i)}$ of each individual's coefficient of agreement; and finally the group coefficient of agreement is the mean of the $\hat{p}_4^{(i)}$ column as indicated by Equation 5, i.e., $\hat{P}_4 = 7.5196/10 \doteq .75$.

As a check on the reasonableness of the estimate above a second aet of X_i' scores, ahown in the last column of Table 1, were generated in the same way as the X_i scorea. Since eight of the students are consistently classified aa master/master or nonmaster/nonmaster on both tests with $C = 4$ (Students 2 and 8 being the exceptions), the two-administration estimate of the coefficient of agreement ia $p_o = 8/10 = .80$. A comparison of the one- and two-administration estimates across criteria $C = 1, 2, 3, 4$ for the example of Table 1 indicatea a median difference of 3 percent between the two indices.

However, the proof of the pudding is in the eating; so let us now consider aome real teat data. In 1974, Form 4B of the Mathematics Baaic Concepts Subteat (Sequential Tests of Educational

Progress--Series II) was administered in Grade 5 of the Madison Public Schools. This is a 50-item multiple-choice test of factual recall, mathematical manipulation, and so forth. A group of $N = 30$ students was selected for analysis, and an odd-item score X_i and an even-item score X_i' were computed for each student, providing unimodal distributions of scores on two, roughly parallel tests of $n = 25$ items. Summary statistics for the unimodal scores X_i and X_i' were as follows: (a) $M_X = 17.40$ and $M_{X'} = 17.17$, (b) $S_X^2 = 5.14$ and $S_{X'}^2 = 4.47$. As in Table 1, a single-administration estimate based on the X_i scores was compared for reasonableness to a corresponding dual-administration estimate based on both X_i and X_i' scores. The results are shown in Figure 1.

Since the distribution of X_i in Figure 2(*) has small variance, as might be expected on a criterion referenced test, the norm referenced reliability coefficient α_{21} is seriously attenuated--$\alpha_{21} = (25/24)[1-(17.40)(25-17.40)/(25x5.14)] \doteq 0$. Thus by Equation 16, $\hat{p}_i = 0(X_i/25) + (1-0)(17.40/25) \doteq .70$ for each of the 30 students. Using the procedure outlined in Table 1 a value of $\hat{P}_C$ was computed for C = 10,11,...,25 as indicated by the broken line in Figure 2. The two-administration estimate p_o (Swaminathan, et al., 1974) based on both X_i and X_i' scores was also computed for the same C values, as indicated by the solid line in Figure 2. The median difference between the two curves is 3 percent across criteria C = 10,11,...,25.

Figure 2 illustrates that $\hat{P}_C$ is a reasonable estimate in the sense that it increases and decreases at points of light and heavy score density (*) in the same way as p_o. However, it would be most unwise to draw conclusions about the accuracy of $\hat{P}_C$ relative to p_o on the basis of this single data set. In this particular case, $\hat{P}_C$ generally provides a conservative estimate of the proportion of consistent decisions relative to p_o. This can be accounted for by two factors: (a) X_i and X_i' are not based on independent administrations as

aasumed by the model for $\hat{P}_C$, so p_o estimates tend to be larger than $\hat{P}_C$ eatimatea; and (b) the aimple binomial model is an approximation to reality. In regard to the latter point, theory auggesta that the compound binomial model of Equationa 12-14. would further enhance the agreement between the curvea of Figure 2.

Generalization to k Mastery Levels

Suppoae there are k posaible maatery levela defined by $k - 1$ criteria $C_1, C_2, \ldots, C_{k-1}$. For example $k = 3$ maatery atates like below-average, average and above-average might be defined by two criterion scores C_1, C_2. Then the probability that person i ia consistently claaaified is given by a general form of Equation 4:

$$
\begin{aligned}
\sim p^{(i)}_{C_1 C_2 \ldots C_{k-1}} &= P(X_i < C_1, X_i' < C_1) + P(C_1 \leq X_i < C_2, C_1 \leq X_i' < C_2) \ldots \\
&\quad + P(C_{k-2} \leq X_i < C_{k-1}, C_{k-2} \leq X_i' < C_{k-1}) + P(C_{k-1} \leq X_i, C_{k-1} \leq X_i') \\
&= [P(X_i < C_1)]^2 + [P(C_1 \leq X_i < C_2)]^2 + \ldots \qquad (19) \\
&\quad + [P(C_{k-2} \leq X_i < C_{k-1})]^2 + [P(C_{k-1} \leq X_i)]^2
\end{aligned}
$$

where the aecond line of Equation 19 again followa from the assumption that X_i and X_i' are independently and identically distributed. If X_i is again aaaumed to have a simple or compound binomial distribution, each term in the bottom line of Equation 19 can be estimated by summing binomial probabilities aa in Equation 8 or by summing compound binomial probabilitiea as in Equation 12. For example, if the simple binomial model is asaumed,

$$P(C_1 \leq X_i < C_2) = \sum_{X_i = C_1}^{C_2 - 1} \binom{n}{X_i} p_i^{X_i} (1 - p_i)^{n - X_i}.$$

The group probability of consistent classification is then obtained by averaging the $P^{(i)}_{C_1C_2\ldots C_{k-1}}$ as in Equation 5:

$$P_{C_1C_2\ldots C_{k-1}} = \frac{\sum_{i=1}^{N} P^{(i)}_{C_1,C_2\ldots C_{k-1}}}{N} \tag{20}$$

Finally, Equation 9 can be written more generally to obtain the group probability of consistent classification due to chance with criteria $C_1, C_2, \ldots, C_{k-1}$ as follows:

$$\begin{aligned} P_{\text{chance}/C_1C_2\ldots C_{k-1}} &= [P(X \leq C_1)]^2 + [P(C_2 \leq X < C_3)]^2 + \ldots \\ &\quad + [P(C_{k-2} \leq X < C_{k-1})]^2 + [P(C_{k-1} \leq X)]^2 \end{aligned} \tag{21}$$

where, for example, $P(C_1 \leq X < C_2)$ is obtained as in Equation 10:

$$P(C_1 \leq X < C_2) = \frac{\sum_{i=1}^{N} P(C_1 \leq X_i < C_2)}{N} \tag{22}$$

Coefficient kappa κ is then obtained as in Equation 11, substituting $P_{C_1C_2\ldots C_{k-1}}$ and $P_{\text{chance}/C_1C_2\ldots C_{k-1}}$ as defined above.

Reference Notes

1. Raju, N. S. A note on Livingston's reliability coefficient for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

2. Harris, C. W. An index of efficiency for fixed-length mastery tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

3. Marshall, J. L. & Haertel, E. H. A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April, 1975.

4. Marshall, J. L. The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data. (Tech. Rep. 350). Madison, WI: University of Wisconsin, Research and Development Center for Cognitive Learning, 1975.

References

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. Psychological Bulletin, 1968, 70, 213-220.

Cohen, J. Weighted chi-square: An extension of the kappa method. Educational and Psychological Measurement, 1972, 32, 61-74.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and methods for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9 27-29.

Harris, C. W. Note on the variances and covariances of three error types. Journal of Educational Measurement, 1973, 10, 49-50.

Livingston, S. A. A reply to Harris's "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." Journal of Educational Measurement, 1972, 9, 31. (a)

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-21. (b)

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability." Journal of Educational Measurement, 1972, 9, 139-140. (c)

Livingston, S. A. A note on the interpretation of the criterion-referenced reliability coefficient. Journal of Educational Measurement, 1973, 10, 311.

Lord, F. M. An Approach to Mental Test Theory. Psychometrika, 1959, 24, 283-302.

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

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

Shavelson, R., Block, J., & Ravitch, M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Swaminathan, H., Hambleton, R. K. & Algina, J. J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Footnotes

The comments of Lawrence Hubert and Joel Levin on an earlier draft of this paper and the assistance of Darwin Kaufman and Barbara Albrecht in obtaining data are gratefully acknowledged.

[1]Listings of a computer program that computes the indices discussed in this paper will be mailed upon request.

Table 1

Estimation of $P_C^{(i)}$ and P_C Using Simulated Data for Ten Persons on a Five Item Mastery Test With Criterion C = 4

i	p_i	X_i^a	$\hat{p}_i^b$	$\hat{P}(X_i \geq 4)$	$1-\hat{P}(X_i \geq 4)$	$[\hat{P}(X_i \geq 4)]^2$	$[1-\hat{P}(X_i \geq 4)]^2$	$\hat{P}_4^{(i)}$	$X_i'^c$
1	.2	0	.19	.0055	.9945	.0000	.9890	.9890	2
2	.4	4	.66	.4478	.5522	.2005	.3049	.5054	2
3	.4	2	.43	.1121	.8879	.0126	.7884	.8010	2
4	.5	0	.19	.0055	.9945	.0000	.9890	.9890	3
5	.5	2	.43	.1121	.8879	.0126	.7884	.8010	2
6	.5	2	.43	.1121	.8879	.0126	.7884	.8010	2
7	.5	1	.31	.0347	.9653	.0012	.9318	.9330	3
8	.6	3	.54	.2415	.7585	.0583	.5753	.6336	4
9	.6	4	.66	.4478	.5522	.2005	.3049	.5054	5
10	.8	5	.77	.6749	.3251	.4555	.1057	.5612	5
								7.5196	

[a] $M_X = 2.30$, $S_X^2 = 2.61$, $\alpha_{21/X} = .58$

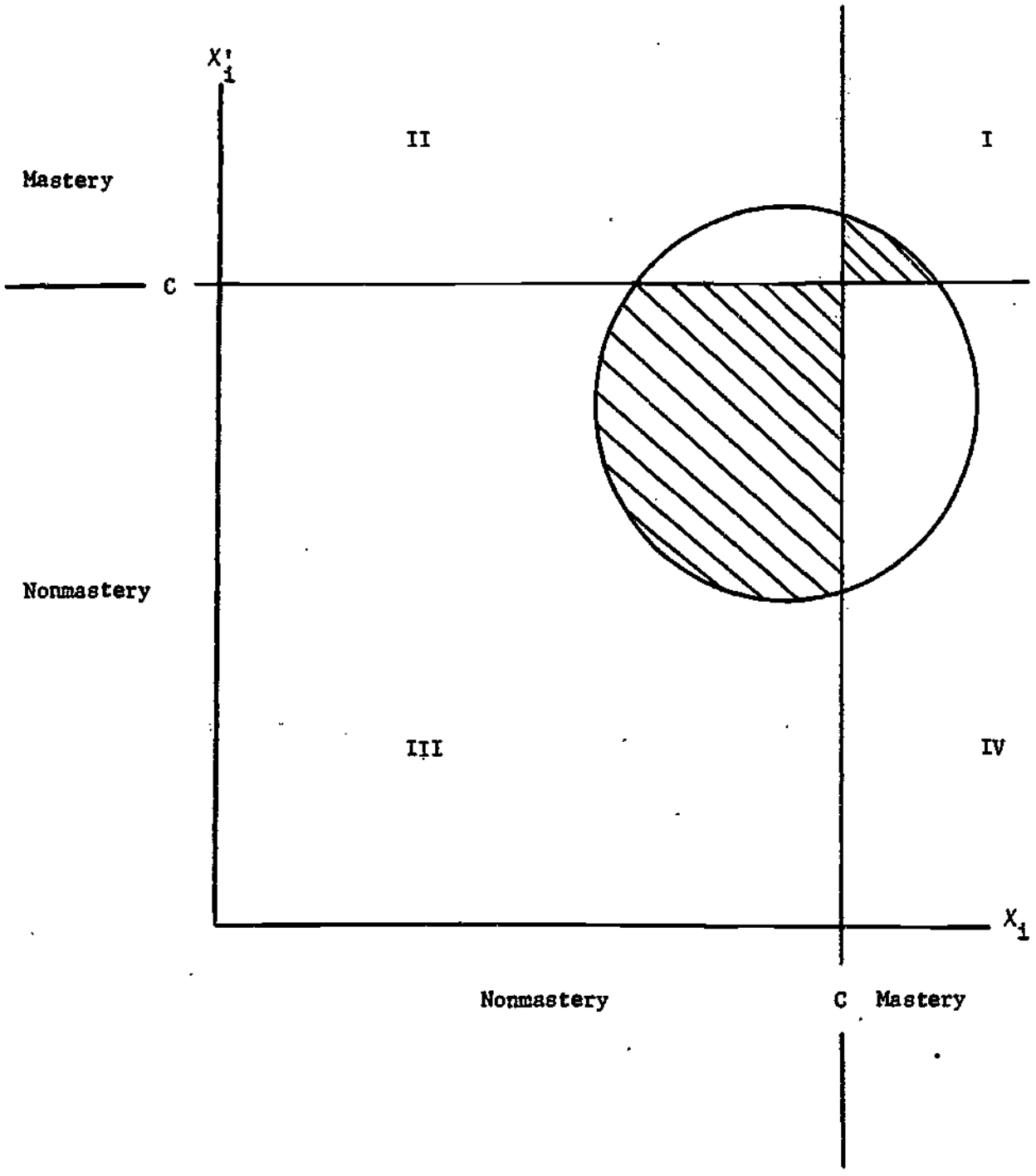
[b] $\hat{p}_i = \alpha_{21/X}(\frac{X_i}{n}) + (1 - \alpha_{21/X})(\frac{M_X}{n})$

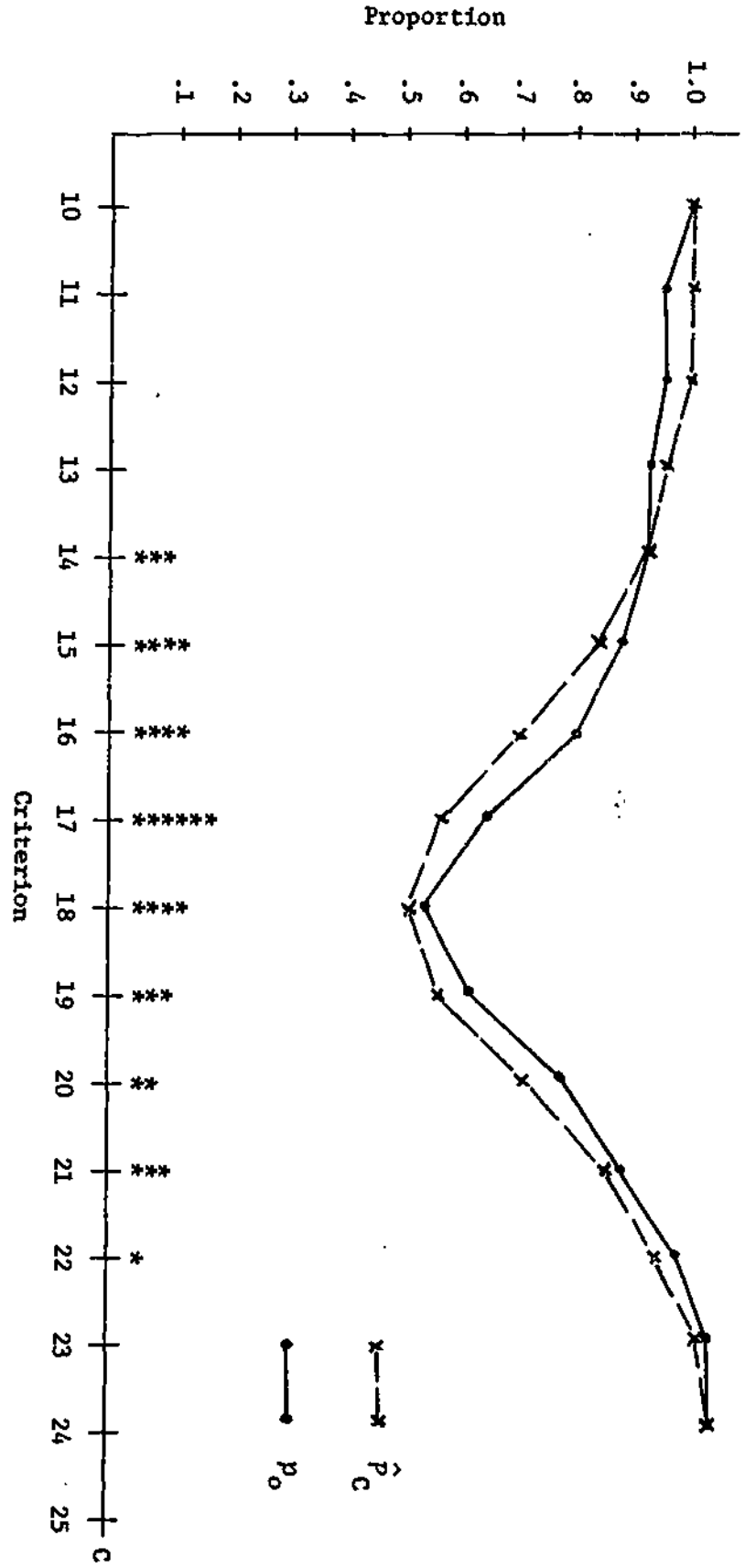
[c] $M_{X'} = 3.00$, $S_{X'}^2 = 1.40$, $\alpha_{21/X'} = .67$

Figure Captions

Figure 1. Outcomes Over Repeated Administrations of Parallel Tests to an Individual

Figure 2. Comparison of One- and Two-Administration Indices For Various Criterion Points in a Unimodal Distribution


X'_i
II
I
Mastery
C
Nonmastery
III
IV
X_i
Nonmastery
C
Mastery


Proportion
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
C
Criterion
***
****
****
******
****
***
**
***
*
p_o
$\hat{p}_c$