DOCUMENT RESUME

ED 171 778 TM 009 239

AUTHOR	Marshall, J. Laird; Serlin, Ronald C.
TITLE	Characteristics of Four Mastery Test Reliability Indices: Influence of Distribution Shape and Cutting Score.
PUB DATE	Apr 79
NOTE	19p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Cutting Scores; *Mastery Tests; *Mathematical Formulas; Mathematical Models; *Scores; *Statistical Studies; *Test Reliability
IDENTIFIERS	*Skewed Distributions

ABSTRACT

Four recent indices emphasizing the interrelationships of score distribution shape, modality, mean, and variance were investigated to determine the reliability of mastery tests. Attention was focused on the values of the indices when the cutoff score was near to or far from the modes of distribution. Five types of score distributions were examined: bell shaped; highly negatively skewed unimodal; bimodal with a stronger mode at the upper end; symmetric bimodal with modes well separated; and symmetric bimodal with modes near each other. Indices examined were based on: (1) Brennan and Kane's index of dependability; (2) Huynh's single administration estimate of the kappa coefficient of agreement; (3) the mean split-half coefficient of agreement, a revision of an earlier formulation by Marshall and Haertel; and (4) Subkoviak's single administration estimate of the coefficient of agreement. (Results are discussed and an explanation of symbols and formulas are appended.) (MH)

Characteristics of Four Mastery Test Reliability Indices:

Influence of Distribution Shape and Cutting Score

J. Laird Marshall

Ronald C. Serlin

University of Wisconsin-Madison

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1979.

Session number 24.27

Purpose

The issue of how to determine mastery test reliability is less than fully settled, particularly when it comes to the issue of which index to use. The purpose of this study was to shed some quasi-empirical light on the subject by examining four relatively recent indices, with attention to the interrelationships of score distribution shape, modality, and proximity of mastery cutoff score to areas of heavy score density (modes). The four single administration indices examined were (sometimes variations or revisions of) those due to Brennan and Kane (19 , Huynh (1976), Marshall and Haertel (1 5), and Subkoviak (1976a).

Many investigators in this field hold that mastery test reliability should deal with the consistency of mastery/nonmastery decisions, or of allocation to mastery state rather than the classical notion of consistency of score itself. Fl ns n an individual's actual score in parallel or repeated testing ms are not considered important, it is claimed, unless they also result in inconsistent mastery state categorizations. Yet, viewing the situation ically, one is hard pressed not to conclude that scores grouped near the should somehow contribute less to the mastery test's reliability than those that are more distant from the cutoff. The attention in this study was focused on the values of those indices when the cutoff score is near or far from the mode(s) of the distribution.

Procedure

A computer program designed and written by the authors, generated item-by-examinee response (correct or incorrect) matrices, according to parameters selected to control score distribution shape, modality, mean, and variance. From each matrix, the score distribution was obtained, and test indices were calculated for all integral cutoff scores. Index value as a function of cutoff score was graphed, as was the relative frequency distribution of scores, so that score distribution shape and mode(s), cutoff score, and index value could be visually compared. The rationale for doing this was that, for a given score distribution shape, index values should be relatively lower when cutoff score is near a mode and relatively higher when cutoff score is in an area of very light score density, if the index is to reflect the property mentioned earlier.[1]

Five types of score distributions were investigated: bell-shaped ( ), highly negatively skewed unimodal ( ), bimodal with a stronger mode at the upper end ( ), symmetric bimodal with modes well separated ( ), and symmetric bimodal with modes near each other ( ). These shapes were only approximately obtained, since the computer program has built-in random error components in order to simulate the results of actual test-taking situations. Each score distribution shape was investigated for tests of 5, 10, and 20 items.

Indices[2]

A. Index of dependability, $M(C)$ (Brennan and Kane, 1977)

Brennan and Kane choose not to call this a reliability index, for reasons discussed in the reference cited. *It* was included in this study, however, in order to see whether it shared any properties with the others. The index is similar to that of Livingston (1972), but is based on generalizability theory rather than classical test theory. The index is defined in terms of expected squared deviations from the cutoff score, C.

B. 1. A single-administration estimate (Huynh, 1976) of the kappa coefficient, $\hat{\kappa}$ (Cohen, 1960).

This estimate assumes that true scores follow a beta distribution with parameters estimated from the mean and variance of the observed score distribution; responses on parallel tests are independent and follow the binomial error model.

2. A single-administration estimate of the coefficient of agreement (proportion of consistent decisions), $\hat{p}$ (Huynh, 1976).

The same assumptions apply here.

C. Another single-administration estimate of the coefficient of agreement, P_C (Subkoviak, 1976).

This index is based on the assumption that the probability that each person is assigned to the same mastery state on parallel tests follows a

1 It is recognized that other criteria may be employed in evaluating reliability indices. In this paper, however, the criterion addressed is whether the indices reflect score distribution mode(s).

2 Appendix A contains all computing formulas used.

binomial form, and incorporates a true score for each person estimated via linear regression using observed score, and observed score mean and variance.

D. The mean split-half coefficient of agreement, β (Marshall, 1976), which is a revision of an earlier formulation (Marshall and Haertel, 1975).

This index is equal to the mean (over persons) proportion (over all possible test splits) of consistent mastery decisions on a hypothetical double-length test, scores on which can be estimated in a variety of ways. Five different methods, or models, for estimating double-length test scores or score distributions were used in this study, and are outlined in Appendix A.

E. In addition, four of the above indices -- κ, p, P_C, β (Huynh, 1978; Subkoviak, 1977; Marshall, 1976) -- can be generalized to multiple mastery states (more than one cutoff score.) These generalized indices were also investigated in this study, but only for the case of three mastery states.

F. Because of the close association between κ and the fourfold correlation index (phi coefficient), ϕ was also calculated on the basis of quantities generated in the calculation of coefficient β, in order that κ might be compared with ϕ for each model.

Results

Although the study generated a great deal of data, the focus reported here is on the degree to which the indices reflect score distribution modes.[3]

1. The index of dependability, $M(C)$, is clearly different from the others, and the conclusion is that it measures quite different things. It did not reflect score distribution modes (except, of course, when the distribution was unimodal and the mean and mode coincided, since, as Brennan and Kane indicate, $M(C)$ always has a minimum at $\overline{X}$.) In fact, $M(C)$ shows the same relationship to *KR21* as Livingstons's index does to *KR20*: the minimum value of each coefficient, occuring at the score distribution mean, equals the respective Kuder-Richardson estimate.

----

3 A more complete report will soon be available and may be obtained by writing either of the authors.

2. Coefficient κ also measures very different things than do p, P_C, and β, as its formulation suggests. Not only did it not reflect score distribution modes (except, by coincidence, when the distribution was bimodal at the extremes), but it behaved in the very opposite way for unimodal distributions, having a maximum rather than a minimum at the mode for symmetric distributions and near the mode for skewed distributions. This is because κ takes on its maximum value in the vicinity of the test mean.

3. Huynh's p did reflect the score modes for unimodal distributions--that is, distributions which approximate one of the beta family, in accordance with the assumptions for that index. The coefficient did not, however, reflect score modes when the shape of the score distribution was bimodal, which is often the case for mastery tests, unless the modes were so extreme as to copy one of the J-shaped or U-shaped beta distributions (a situation which is not likely to happen in the real world, particularly when guessing occurs). Based on the research of this study, the authors hypothesize that the p coefficient would fare better on this criterion if Huynh had chosen a predictive Bayesian beta-binomial approach (Aitchison and Dunsmore, 1975), akin to D. I. in Appendix A, even though that approach is slightly more complex. Although earlier research (Subkoviak, 1978) recommended the Huynh procedure, it should be noted that Subkoviak's study dealt only with unimodal distributions closely approximating a beta distribution. It is likely that the recommendations would have been otherwise had bimodal distributions similar to those in this study been investigated.

4. Subkoviak's P_C generally reflected score modes very well, for both unimodal and bimodal distributions. The one exception was when the distribution was bimodal and the modes were close together, particularly for short tests.[4] But since this type of score distribution is atypical, the Subkoviak approach is, overall, highly satisfactory.

5. Of the five estimation models for Marshall's coefficient β, model 1 was the least satisfactory, for reasons to be discussed in part 7 of this section.

4 In this situation, a compound rather than a simple binomial model would fare better; in all other cases the simple and compound binomial models yielded nearly identical results, supporting Subkoviak's (1978) findings.

Model 1 reflected score modes, but uniformly not as well as did models 2, 4, and 5, except when the distribution was unimodal. Models 2 and 4 were nearly identical; the only exception being the situation described above for Subkoviak's P_C. Model 4 is thus the preferable of the two for reasons of simplicity. Thus the choice narrows down to models 4 and 5. Model 5 yielded better results in the situation described above, and slightly better results when the two modes are widely separated. Model 4 yielded better results for short tests for the asymmetric bimodal case. Other than that, the two models were comparable and yielded very satisfactory results vis-a-vis the mode reflection criterion.
6. For the three-mastery-state indices, a trimodal distribution was constructed; p, P_C, and β model 2, 3, 4, and 5 were calculated for various combinations of cutoff scores. Of these, p and β_3 did not reflect score modes. P_C and the other three models did; interpretation is more difficult, however, since the graphs involved should be 3-dimensional (index value as a function of the two cutoff scores), and the computer program was not set up to handle this situation. The authors plan to research this topic further.
7. This study produced another significant finding, which might have been but was not deduced mathematically beforehand, and thus rendered an element of surprise. Although the following results have not yet been proved rigorously (the authors are working on it), the computer-calculated empirical evidence is so overwhelming that we feel secure in claiming the following conjectures:

i) Since p requires assumptions about a beta-binomial distribution, if analogous assumptions (for a double- rather than single-length test) are postulated for coefficient β, model 3, the two indices are identical, i.e., $p = \beta_3$. This explains why coefficient β, model 3, had unsatisfactory characteristics. This conjecture (and the two to follow) is backed up by over 300 pairs of calculated index values indentical to three decimal places, over all ranges of cutoff score, distribution type, and test length. Moreover, one would suspect that if we used the predictive Bayesian formulation as suggested earlier, it would under these conditions equal β_1.

ii) Since κ requires the same assumption as does p, when the phi coefficient is calculated according to the formula in Appendix A under model 3, $\kappa=\phi$.

iii) Since P_C entails assumptions about binomial error and a regression estimate of true score, if analogous (for $2n$ items) assumptions are made for coefficient β, model 4, the two indices are identical, i.e., $P_C = \beta$. It is further hypothesized that if the compound binomial model were used for P_C, then under these conditions P_C would equal β.

It is apparent, then, that the question of whether to employ the Huynh p or Subkovial P_C or Marshall p is not relevant, since each of the first two is a special instance of the third more general coefficient, when the appropriate assumptions are answered. The question instead should be which set of assumptions is appropriate for the situation.

References

Aitchison, J. & Dunsmore, I. R. Statistical Prediction Analysis. Cambridge: Cambridge University Press, 1975.

Brennan, R. L. & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, pp. [illegible]-289.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 213-220.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, pp. 253-264.

Huynh, H. Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Paper presented at the [illegible]7th Annual Southeastern Invitational Conference on Measurement in Education, Greensboro, South Carolina, December, 1978.

Livingston, S. A. A criterion-referenced application of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, pp. 239-269.

Marshall, J. L. The mean split-half coefficient of agreement and its relation to other test indices: a study based on simulated data. Technical Report No. 350, Madison: Wisconsin Research and Development Center for Cognitive Learning, 1976.

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., 1975.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, pp. 265-276.

Subkoviak, M. J. Further comments on reliability for mastery tests (Laboratory of Experimental Design, Occasional Paper No. 17). Unpublished manuscript, University of Wisconsin, 1977.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, pp. 111-116.

Wilcox, Rand R. Estimating true score in the compound binomial error model. Psychometrika, 1978, 43, pp. 245-258.

Appendix A : Symbols and Formulas

Different authors use different symbols for the same thing. In order to minimize confusion, we have in this paper used a set of symbols that are as close as feasible to the authors' originals yet which are in common usage and have the same meaning throughout. If this is a compromise, we hope it is a compromise in favor of consistency and clarity.

Symbols

In what follows, these symbols have a common meaning:

n = number of test items

N = number of persons

x = an obtained test score, $0 \leq x \leq n$

$f(x)$ = frequency of score x in the obtained score distribution

$\overline{X}$ = test mean

S_x^2 = score variance

C = mastery cutoff score (where $X \geq C$ denotes "mastery"), $0 < C \leq n$

α_{21} = Kuder-Richardson formula 21

α_{20} = Kuder-Richardson formula 20

Computing Formulas

A. $M(C)$ (Brennan & Kane, 1977)

$$M(C) = 1 - \frac{1}{n-1}\left[\frac{\overline{X}(n - \overline{X}) - S_x^2}{(\overline{X} - C)^2 + S_x^2}\right]$$

B. 1. κ (Huynh, 1976; Cohen, 1960)

$$\kappa = (p_{11} - p_1^2)/(p_1 - p_1^2)$$

where $p_{11} = \sum_{x,x'=C}^{n} h(x,x')$

and $p_1 = \sum_{x=C}^{n} h(x)$

Here, $h(x)$ is the univariate negative hypergeometric density, $\binom{n}{x}B(a+x,n+b-x)/B(a,b)$ and $h(x,x')$ is the bivariate density, $\binom{n}{x}\binom{n}{x'}B(a+x+x',2n+b-x-x')/B(a,b)$ in which B represents the beta function and a and b are parameters estimated by

$$a = \overline{X}\left(\frac{1-\alpha_{21}}{\alpha_{21}}\right)$$

$$b = (n-\overline{X})\left(\frac{1-\alpha_{21}}{\alpha_{21}}\right)$$

2. $\underline{p}$ (Huynh, 1978)

$$p = \sum_{x,x'=0}^{C-1} h(x,x') + \sum_{x,x'=C}^{n} h(x,x')$$

where $h(x,x')$ is as previously defined.

C. $\underline{P_C}$ (Subkoviak, 1976)

$$P_C = \frac{1}{N}\sum_{x=0}^{n} f(x)\left(\left[P(X\geq C)\right]^2 + \left[1-P(X\geq C)\right]^2\right)$$

in which $P(X\geq C) = \sum_{j=C}^{n}\binom{n}{j}\theta_x^{\,j}(1-\theta_x)^{n-j}$

where θ represents the true score of a person with obtained score of x and is estimated by

$$\theta_x = \alpha_{21}\left(\frac{x}{n}\right) + (1-\alpha_{21})\left(\frac{\overline{X}}{n}\right)$$

D. <u>β</u> (Marshall, 1976)

In what follows,

y = a possible score on a hypothetical double-length test, $0 \leq y \leq 2n$

$f(y)$ = estimated frequency of score y in the (hypothecial) double-length test score distribution

$$\beta = \frac{1}{N}\left[\sum_{y=0}^{C-1} f(y) + \sum_{y=C}^{2C-2} f(y)\cdot H_y(y-[C-1],C-1) + \sum_{y=2C}^{n+C-1} f(y)\cdot H_y(C,y-C) + \sum_{y=n+C}^{2n} f(y)\right]$$

where $H_y(\ell,m)$ is a partial sum of hypergeometric terms:

$$H_y(\ell,m) = \sum_{j=\ell}^{m} \frac{\binom{y}{j}\binom{2n-y}{n-j}}{\binom{2n}{n}}$$

Note that the second term in the brackets above vanishes when C=1 and the third term vanishes when $C=n$.

In this, $f(y)$ can be estimated in a number of ways. We have chosen five estimation techniques which correspond with the different models for coefficient β discussed in the presentation.

<u>1</u>. Predictive Bayesian beta (Aitchison & Dunsmore,1975)

$$f(y) = \sum_{x=0}^{n} f(x) \cdot\binom{2n}{y} B(a+x+y, 3n+b-x-y)/B(a+x,n+b-x)$$

where B is the beta function and a,b are estimated as in the Huynh procedure.

2. Compound binomial (Lord, 1965; adjusted for $2n$ items)

$$f(y) = \sum_{x=0}^{n} f(x) \cdot b(y;2n,\theta_x) \cdot \left[1 + \frac{kQ}{2n(2n-1)\theta_x(1-\theta_x)}\right]$$

where $b(y;2n,\theta_x) = \binom{2n}{y} \theta_x^{\,y} (1-\theta_x)^{2n-y}$

$$\theta_x = \alpha_{20}\left(\frac{x}{n}\right) + (1-\alpha_{20})\left(\frac{\overline{X}}{n}\right)$$

$$Q = -2n(2n-1)\theta_x^{\,2} + 2y(2n-1)\theta_x - y(y-1)$$

$$k = \frac{n^2(n-1)S_\pi^{\,2}}{\overline{X}\,(n-\overline{X})-S_x^{\,2}-nS_\pi^{\,2}}$$

in which $S_\pi^{\,2}$ is the variance of the item difficulties.

3. Beta distribution with parameters that are functions of the obtained score distribution(similar to that used for Huynh's coefficients, but adjusted for $2n$ items).

$$f(y) = N\binom{2n}{y} B(a+y,\ 2n+b-y)/B(a,b)$$

where B, a, b are as before.

4. Binomial Regression (similar to that used by Subkoviak in his index, but adjusted for $2n$ items)

$$f(y) = \sum_{x=0}^{n} f(x) \cdot \binom{2n}{y} \theta_x^{\,y} (1-\theta_x)^{2n-y}$$

where θ_x is as in Subkoviak's coefficient.

5. Averaged "double binomial"

This one was conjured up by the authors in an attempt to find an $f(y)$ estimate that does a better job than do most of the others in echoing the modes of the obtained X distribution. Although mathematically less defensible, its empirical properties are generally good.

$$f(y) = n\binom{2n}{y}\left\{2f(0)B_{\frac{1}{2n}}(1,2n+1) + \sum_{x=1}^{n-1} f(x)\left[B_{\frac{(x+\frac{1}{2})}{n}}(1,2n+1) - B_{\frac{(x-\frac{1}{2})}{n}}(1,2n+1)\right] + 2f(n)\left[1 - B_{1-\frac{1}{2n}}(1,2n+1)\right]\right\}$$

where $B_v(r,s) = \int_0^v t^{r-1}(1-t)^{s-1}dt$ and $t = \frac{x}{n}$.

Other models were considered but rejected either as too complex and not worth the trouble or on the grounds that preliminary research showed them to have undesirable characteristics. The former category includes various methods which involve arcsine transformations and some methods suggested by Wilcox (1978) (to whom we acknowledge appreciation for suggesting the Aitchison & Dunsmore reference); the latter category includes the unbiased $\left(\frac{x}{n}\right)$ estimate of θ_x and a predictive Bayesian model with vague priors. Preliminary research also showed that model 3 outlined above fell in this category, but it was retained for comparison purposes.

E. Indices which allow multiple mastery states. For this study, no more than three mastery states were considered; the following formulas have been simplified accordingly. In these formulas, K is the lower and C is the upper cutoff score, $0<K<C\leq n$.

1. $\kappa(K,C) = (p.-p_*)/(1-p_*)$ (Huynh, 1978)

   where $p. = \sum_{x,x'=0}^{K-1} h(x,x') + \sum_{x,x'=K}^{C-1} h(x,x') + \sum_{x,x'=C}^{n} h(x, x')$

   and $p_* = \left[\sum_{x=0}^{K-1} h(x)\right]^2 + \left[\sum_{x=K}^{C-1} h(x)\right]^2 + \left[\sum_{x=C}^{n} h(x)\right]^2$

   where h is as defined for κ.

2. $p(K,C) = p.$ as defined above. (Huynh, 1978)

3. $P(K,C)$ (Subkoviak, 1977)

   $$P(K,C) = \frac{1}{N}\sum_{x=0}^{n} f(x) \cdot \left[(m_{Kx})^2 + (m_{Cx})^2 + (1-m_{Kx}-m_{Cx})^2\right]$$

   where $m_{Kx} = \sum_{j=K}^{C-1} \binom{n}{j} \theta_x^{\,j} (1-\theta_x)^{n-j}$

   $m_{Cx} = \sum_{j=C}^{n} \binom{n}{j} \theta_x^{\,j} (1-\theta_x)^{n-j}$

   and θ_x is as defined for P_C.

4. $\beta(K,C)$ (Marshall, 1976)

$$\beta(K,C) = \frac{1}{N}\left[\sum_{y=0}^{K-1} f(y) + \sum_{y=K}^{2K-2} f(y)\cdot H_y(y-[K-1],K-1) + \sum_{y=2K}^{2C-2} f(y)\cdot H_y(u,v) + \sum_{y=2C}^{n+C-1} f(y)\cdot H_y(c,y-c) + \sum_{y=n+c}^{2n} f(y)\right]$$

where $f(y)$ depends on the model,

$u = max\ (K, y-[C-1])$

$v = min\ (C-1,\ y-K)$

and $H_y(\ell,m)$ is as defined for β.

Note that the second term in the brackets vanishes when $K=1$ and the fourth term vanishes when $C=n$.

F. Calculation of Phi Coefficient

$$\phi = \frac{AD - E^2}{(A+E)(D+E)}$$

where $A = \sum_{y=2C}^{n+C-1} f(y)\cdot H_y(C,y-C) + \sum_{y=n+C}^{2n} f(y)$

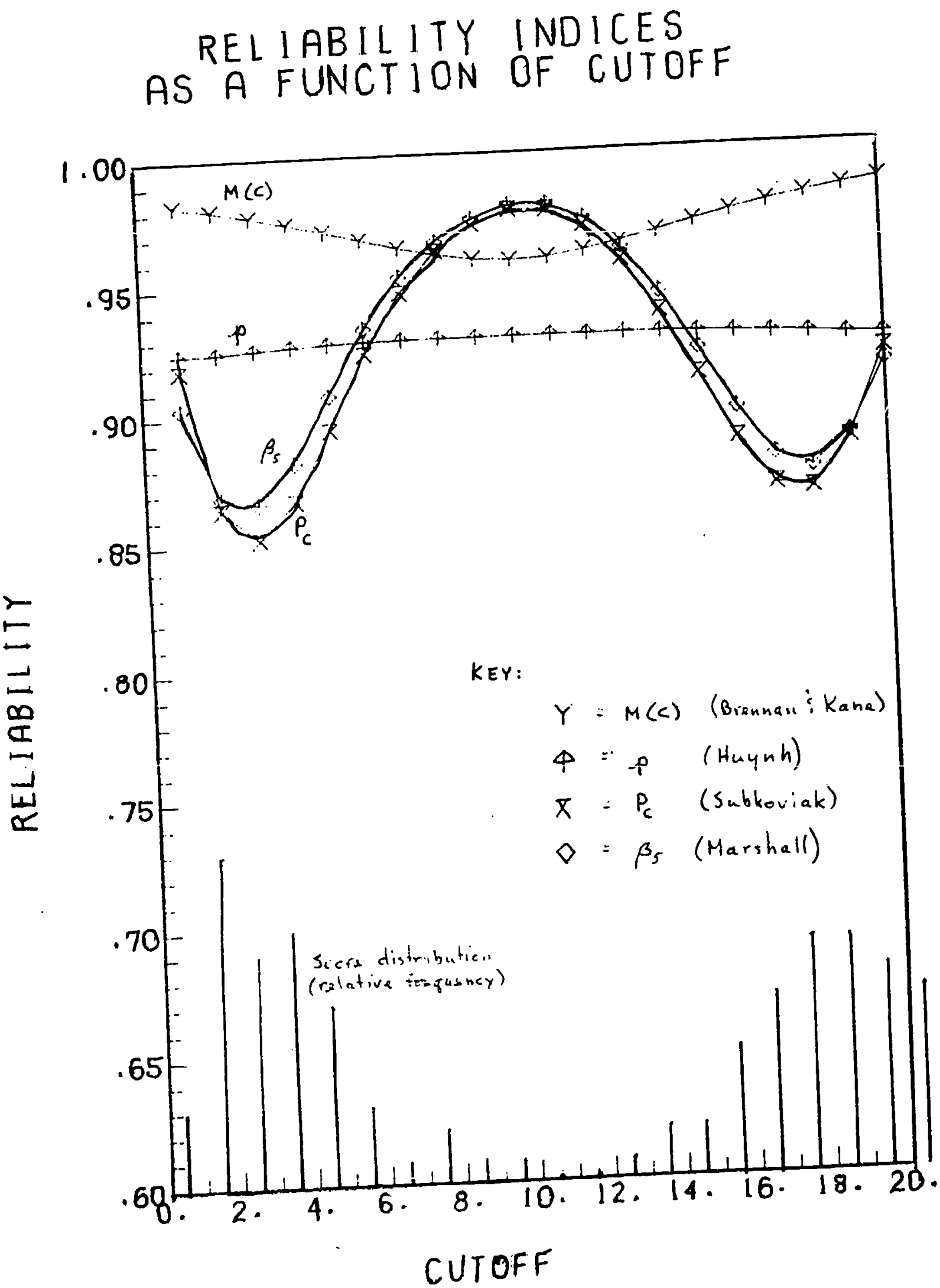
$D = \sum_{y=0}^{C-1} f(y) + \sum_{y=C}^{2C-2} f(y)\cdot H_y(y-[C-1],C-1)$

$E = (N-A-D)/2$

The above expressions for A and D can be seen to be derived from the formula for coefficient β.

Appendix B

Example of a Graph



Appendix C

Y	$M(C)$	(Brennan & Kane, 1977)
Z	κ	(Huynh, 1976)
↑	p	(Huynh, 1976)
X̄	P_C	(Subkoviak, 1976)
⬡	β model 1	Variations of β (Marshall, 1976)
△	β model 2	Variations of β (Marshall, 1976)
+	β model 3	Variations of β (Marshall, 1976)
×	β model 4	Variations of β (Marshall, 1976)
◇	β model 5	Variations of β (Marshall, 1976)
⧗	φ model 3	