Group Dependence of Some Reliability Indices for Mastery Tests

D. R. Divgi
Syracuse University

Reliability indices for mastery tests depend not only on true-score variance but also on mean and cutoff scores. This dependence was examined in the case of three decision-theoretic indices: (1) the coefficient of agreement; (2) kappa; and (3) the proportion of correct decisions. The binomial error model was assumed, with a two-parameter beta distribution for true scores. The reliability indices were computed at five values of the mean, four values of KR-21, and four cutoff scores. Results show that the dependence of kappa on mean and cutoff scores is opposite to that of the proportion of correct decisions, which is linearly related to average threshold loss. Moreover, kappa can be very small when most examinees are classified correctly. Thus, objections against the classical reliability coefficient apply even more strongly to kappa.

The purpose of a mastery test is to classify an examinee into one of two groups—Masters and Nonmasters—by comparing the test score with a criterion value C. Scores of different individuals are not compared and, therefore, score variability is not considered important. A test may provide reliable classification and yet have small score variance, and thus a low classical reliability (Popham & Husek, 1969). This possibility, and the fact that test scores are used in a different way, make it desirable to use other reliability indices for mastery tests.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 2 Spring 1980 pp. 213–218

Review of Reliability Indices

Livingston (1972) proposed a modification of the classical coefficient of reliability:

$$k^2 = \frac{\sigma^2(\zeta) + (\mu - C)^2}{\sigma^2(X) + (\mu - C)^2} \qquad [1]$$

where ξ and X are true and observed scores and μ is the mean score. Brennan and Kane (1977) have used generalizability theory to derive an index of dependability, which equals k^2 when item difficulties are equal. Another index based on the concepts of analysis of variance has been proposed by Harris (1974).

In order to compute any of these indices, the actual score obtained by each examinee is needed, and not just the classification. Instead, one may take the decision-theoretic viewpoint that the examinee's classification is the only outcome of interest, and no further attention need be paid to the score. One index based on this viewpoint is the coefficient of agreement p_a, the proportion of examinees who obtain the same classification in two parallel sets of measurements on the same group (Hambleton & Novick, 1973; Subkoviak, 1976). Another index is Cohen's (1960) kappa, which was recommended by Swaminathan, Hambleton, and Algina (1974) and has been studied extensively by Huynh (1976, 1977, 1978). It, too, is based on consistency between two sets of decisions, but with one

important difference. Kappa measures the improvement the test provides over chance agreement, expressed as a fraction of the maximum possible improvement over chance.

Let p_1 and p_0 be the proportions of individuals classified as Masters and Nonmasters (averaged over the two measurements). Then, the coefficient of agreement due to pure chance is

$$p_{ch} = p_1^2 + p_0^2 \qquad [2]$$

and kappa is given by

$$\kappa = (p_a - p_{ch})/(1 - p_{ch}). \qquad [3]$$

The smallest value of p_{ch} is .5, which occurs when $p_1 = p_0 = .5$. Therefore, p_a is never less than .5. Kappa, however, can vanish (Huynh, 1978). Both these coefficients can be calculated by using two forms of the test or from a single test administration coupled with a theoretical model. (For a comparison of these methods, see Huynh, 1976, for kappa and Subkoviak, 1976, for the coefficient of agreement.) Marshall and Haertel (1975) have suggested averaging p_a over all possible split halves. This, however, requires a great deal of computation.

Another index describing the accuracy of decisions is the proportion of correct decisions, where for an individual the correct decision is defined by comparing the true score with the criterion. This index has been considered by Subkoviak and Wilcox (1978). It will be denoted by p_c.

Purpose of the Study

Values of all the reliability indices mentioned above depend on the distribution of true scores in the population, in particular on the mean and the standard deviation. This makes it difficult to interpret these indices and to draw conclusions about relative merits of different tests. If the reliability quoted is higher for one test than for another, the cause may lie in the tests or in the populations to which they were administered. In the norm-referenced case, if necessary data are available, reliabilities can be translated into standard errors of measurement, which are known to be quite stable from one population to another (Lord, 1959). This is not possible with mastery tests because reliability depends not only on heterogeneity of the group but also on the mean score. Therefore, a quoted value of a reliability coefficient cannot be interpreted properly unless the mean and the variance of the scores are provided also.

In order to make use of such additional information, it is necessary to know how population characteristics affect various coefficients. So far, no one has reported a systematic study of this question, with the two relevant quantities varied independently. Huynh (1976) and Subkoviak (1976) reported variations of kappa and p_a, respectively, when the cutoff score was varied for a given distribution of scores. Although this was unavoidable while using real data, it has little relevance to actual practice. The real life situation is one in which groups of different ability levels take the same test with a specified cutoff score. Huynh (1976) has noted that kappa increases with score variance, but his conclusion is based only on single administrations of three different tests.

Another question, perhaps more important than that of interpretation, is this: If two coefficients vary with a population parameter in opposite ways, which coefficient should be used? It is known that kappa is largest when the cutoff C equals the mean score μ (Huynh, 1976) and that the coefficient of agreement is largest when C and μ are far apart (Subkoviak, 1976). In which case are the mastery classifications more "reliable"? Depending on the answer, one must decide to use one coefficient and to reject the other. It may be argued that in practice the cutoff is not varied and therefore that there need be no worry about this contrast. If, however, the same contrast is found when μ is varied, concern is with variation of ability from one group to another, which does occur in practice. In that case, the need to choose between these reliability coefficients cannot be avoided.

Once a coefficient is selected, results similar to those in Table 1 can be used for approximate conversion of the value of the coefficient from one set of conditions to another. Simulations meant for such practical use will have to assume (1) test length equal to that of the actual test and (2) a more realistic model, e.g., the compound binomial. The present study is intended mainly to illustrate trends.

Method

The present study was implemented to determine the effects of population parameters on three indices based on decision theory: (1) the coefficient of agreement p_a, (2) kappa, and (3) the proportion of correct decisions p_c. These were calculated for entire populations rather than for finite samples in order to avoid any blurring of relationships due to sampling errors.

For any given (proportion-correct) true score ζ, the distribution of observed scores was assumed to be binomial. The true score was assumed to have a two-parameter beta distribution (Lord & Novick, 1968, chap. 23). This model has been used frequently in studies of reliability indices for mastery tests (Huynh, 1976, 1977; Subkoviak & Wilcox, 1978). The compound binomial error model can be used instead of the binomial, but the change in results is small. More important, the pattern of the relationship between reliability and an independent variable does not change (Subkoviak, 1976, Figure 1).

The parameters of interest are the mean μ and the variance of true scores. The latter, however, is inconvenient; its maximum possible value depends on the value of the mean, since $0 \leqslant \zeta \leqslant 1$. In contrast, the classical coefficient of reliability can have any value from 0 to 1 at any mean score. In a binomial model the reliability equals α_{21}, given by the Kuder-Richardson Formula 21. Thus, the independent parameters were α_{21} and the mean score μ (which, in the total population, is the same for true and observed scores).

A 10-item test was considered, with four values of the (proportion-correct) cutoff score C = .55, .65, .75, and .85. The values of the mean score and reliability were μ = .45, .55, .65, .75, and .85 and α_{21} = .25, .40, .55, and .70. Selection of the highest and lowest values of the parameters was guided by the values in the real data used by Huynh (1976) and by Subkoviak (1978). Calculations were carried out independently for each set of values of the three independent parameters.

For any individual with true score ζ the probability of classification as Master is $P(x \geqslant C|\zeta)$. In the binomial model this probability is given by the binomial distribution, with ζ being the probability of answering any given item correctly. Its average over the distribution of ζ equals the marginal proportion of examinees classified as Masters, denoted by p_1. This has the same value for test and retest. The proportion of persons classified as Masters in both tests is the expectation of $[P(x \geqslant C|\zeta)]^2$, denoted by p_{11}. Then, the coefficient of agreement p_a can be calculated as $1+2(p_{11} - p_1)$. The value of p_a expected from pure chance is

$$p_{ch} = p_1^2 + (1 - p_1)^2. \qquad [4]$$

Kappa then is given by Equation 3. The probability of correct classification for an individual is $P(x \geqslant C|\zeta)$ if $\zeta \geqslant C$, and $P(x < C|\zeta)$ otherwise. Averaging these yields p_c, the proportion of correct classifications in the population.

As mentioned above, the probability density $g(\zeta)$ at true score ζ was assumed to be a two-parameter beta distribution, i.e.,

$$g(\zeta) = \frac{\zeta^{a-1}(1 - \zeta)^{b-n}}{B(a, b - n + 1)} \qquad [5]$$

where B is the beta function and n is the number of items in the test. The parrameters of the distribution are determined by the mean and the reliability as follows:

$$a = (-1 + 1/\alpha_{21})\, n \qquad [6]$$

$$b = -a - 1 + n/\alpha_{21} \qquad [7]$$

(Lord & Novick, 1968, pp. 517 & 520). Integrals over the beta distribution were evaluated by 24-point Gaussian quadrature.

Results and Discussion

Kappa can vary from 0 to 1, whereas the coefficient of agreement has a non-zero minimum value, which is .5 in binary classifications. This can lead to incorrect impressions while comparing their properties and has been avoided by defining rescaled values through the following linear transformations.

$$p'_a = 2\,p_a - 1, \qquad [8]$$

$$p'_c = 2\,p_c - 1. \qquad [9]$$

The results are shown in Table 1, which displays p'_a and p'_c rather than p_a and p_c. Values of the reliability indices have been shown without decimal points. With some exceptions, all three indices increased when α_{21} increased, while the mean was constant. The exceptions occur in cases of p_a and p_c when $|\mu - C|$ was large. Consider $C = .55$ and $\mu = .85$. The four values of α_{21} correspond to true-score standard deviations of .06, .09, .12, and .16, respectively. These are appreciably smaller than the difference between the mean and the cutoff score. As classical reliability increased, the distribution of the true scores spread out and the probability density near C increased. This raised the proportion of incorrect and inconsistent decisions, because errors of classification are most likely for persons with true score near the cutoff.

At any given value of α_{21}, kappa is largest when μ and C are equal, whether the mean is varied while holding the cutoff score constant (Table 1) or vice versa (Huynh, 1976). In general, p_a and p_c are smallest under these circumstances (see Subkoviak, 1976). This is to be expected, since equality of μ and C implies that the mode of the true score distribution is close to the cutoff value, where errors are most likely. Both p_a and p_c depend on the independent variables in the same way. Indeed, p'_a is approximated by $.35p'_c + .65p'^2_c$ with maximum error less than .03 over a range from .17 to .96. This is based on a wide range of parameter values, so it is unlikely that any other set of realistic μ and α_{21} will yield an appreciably different result (for a 10-item test). The values .35 and .65 are not important; nor is even the functional form, except that it is monotonic. The point to note is the surprisingly small size of the scatter, which means that there is nearly an exact mathematical relationship between consistency and the proportion of correct decisions. Therefore, the two coefficients are practically equivalent, and almost the same information is obtained no matter which one is used.

The literature does not contain any set of criteria for choosing among different indices of reliability. Therefore, consider two arguments that have been leveled against using the classical reliability coefficient for mastery tests. One objection concerns loss functions. According to Hambleton and Novick (1973), the classical coefficient is inappropriate because it is based on a squared-error loss function. They prefer the threshold loss function, with loss in an individual case equal to $a>0$ if the decision is wrong and equal to zero if the decision is correct. Clearly, expected threshold loss is $a(1-p_c)$. Thus, of the three indices considered here, the proportion of correct decisions is related directly to threshold loss. The coefficient of agreement, which was recommended by Hambleton and Novick (1973), is almost as good because its values are closely related to those of p_c. Not so with kappa. In fact, the dependence of kappa on the mean and criterion scores is opposite to that of p_c. Its smallest value in Table 1 is .03, which occurs when $p'_c = .98$.

The other objection against the classical reliability coefficient concerns the dependence of this coefficient on true-score variance in the group. If $|\mu - C|$ is appreciably larger than the standard error of measurement and the true-score variance is small, the test will provide reliable classifications, but its classical reliability may be low (Millman & Popham, 1974; Popham, 1978, p. 144; Popham & Husek, 1969).

Table 1
Decision-Theoretic Reliabilities as Functions of Cutoff Score, Mean Score, and KR-21[a]

Mean Score	C = 0.55 KR-21				C = 0.65 KR-21				C = 0.75 KR-21				C = 0.85 KR-21			
	25	40	55	70	25	40	55	70	25	40	55	70	25	40	55	70
Rescaled coefficient of agreement																
0.45	31	37	45	55	59	58	59	62	83	79	75	72	96	93	89	83
0.55	17	28	39	52	29	36	44	54	59	57	57	61	85	81	76	72
0.65	33	39	47	57	17	28	40	53	29	35	43	53	62	59	58	59
0.75	66	65	66	69	37	43	51	61	18	28	40	54	30	35	42	52
0.85	92	88	85	83	75	73	73	75	46	51	57	66	19	30	42	56
Kappa																
0.45	16	26	38	52	12	22	35	49	08	16	28	44	03	09	20	37
0.55	17	28	39	52	15	26	38	51	12	21	34	48	06	14	26	42
0.65	16	27	38	52	17	28	39	53	15	26	37	51	10	20	32	46
0.75	12	23	35	50	16	27	39	52	17	28	39	53	15	25	37	50
0.85	07	17	31	47	12	23	36	51	16	27	39	53	17	28	40	53
Rescaled proportion of correct decisions																
0.45	49	53	60	68	73	71	70	73	90	87	83	80	98	96	93	88
0.55	34	45	55	65	47	52	59	67	74	70	70	72	92	88	84	80
0.65	51	55	62	69	34	45	55	66	47	51	58	66	76	72	70	71
0.75	79	76	76	78	55	59	64	72	35	46	56	66	48	51	57	65
0.85	95	93	90	89	85	82	81	84	62	65	69	77	36	47	57	70

[a] All coefficients, including KR-21, are shown without decimal points.

This can happen when the group taking the test has been well trained, and use of the classical reliability coefficient may lead to unjustified rejection of the test as unreliable.

The values in Table 1 show that kappa is always smaller than α_{21} and p'_c (which have the same range as kappa) and smaller than p'_a except when $\mu = C$. Assuming a normal distribution of scores, Huynh (1978) has shown that the maximum value of kappa is $(2/\pi)$ arcsin (ρ) where ρ is the classical reliability. This upper limit, in turn, is smaller than ρ unless $\rho = 1$ (Lord & Novick, 1968, Table 15.9.1). The values of ρ used in the present study are .25, .40, .55, and .70. The corresponding values of $(2/\pi)$ arcsin (ρ) are .161, .262, .371, and .494, respectively. The largest values of kappa in Table 1 are larger than these, but only slightly. Thus, we have the double inequality.

$$\kappa \lesssim (2/\pi) \arcsin(\rho) < \rho. \qquad [10]$$

Since kappa is correlational in nature (Huynh, 1976, p. 260), it is sensitive to score variance, just as ρ is. Thus, if the classical coefficient is to be considered unsatisfactory because it can be small even when the accuracy of classification is adequate, the criticism applies even more strongly to kappa.

References

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. *Journal of Educational Measurement,* 1977, *14,* 277-289.

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

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

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement.* Los Angeles: UCLA, Center for the Study of Evaluation, 1974.

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

Huynh, H. *The kappamax consistency index for decisions in domain-referenced testing.* Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Huynh, H. Reliability of multiple classifications. *Psychometrika,* 1978, *43,* 317-325.

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

Lord, F. M. Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement,* 1959, *19,* 233-239.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

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, DC, April 1975.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement,* 1974, *11,* 137-138.

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

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

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

Subkoviak, M. J., & Wilcox, R. *Estimating the probability of correct classification in mastery testing.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

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

Acknowledgments

I am grateful to Eric Gardner for advice on improving the presentation.

Author's Address

Send requests for reprints or further information to D. R. Divgi, 353 Lindquist Center, University of Iowa, Iowa City, IA 52242.