DOCUMENT RESUME

ED 154 004 TM 007 000

AUTHOR	Huynh, Huynh
TITLE	The Kappamax Reliability Index for Decisions in Domain-Referenced Testing.
PUB DATE	Mar 77
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; *Mathematical Models; Statistical Analysis; *Test Reliability
IDENTIFIERS	*Domain Referenced Tests; *Kappamax Reliability Index; Kappa Reliability Index

ABSTRACT

The kappamax reliability index of domain-referenced tests is defined as the upper bound of kappa when all possible cutoff scores are considered. Computational procedures for kappamax are described, as well as its approximation for long tests, based on Kuder-Richardson formula 21. The sampling error of kappamax, and the effects of test length and of score variability on kappamax are discussed. Estimates of kappamax for double-length tests are shown to be related to Spearman-Brown estimates. (CTM)

THE KAPPAMAX RELIABILITY INDEX FOR DECISIONS IN DOMAIN-REFERENCED TESTING

Huynh Huynh
University of South Carolina

Abstract

Within the beta-binomial framework for domain-referenced testing, the kappamax reliability index is proposed for binary decisions based on test scores, and studied in terms of variation, approximation and sampling fluctuation. A formula is given for the projection of kappamax of tests with doubled lengths.

March 1977

This paper will be presented at the annual meeting of the Ameirican Educational Research Association, New York City, April 4-8, 1977.

THE KAPPAMAX RELIABILITY INDEX FOR DECISIONS IN DOMAIN-REFERENCED TESTING

The search for a suitable reliability index for criterion-referenced tests (Hambleton & Novick, 1973) seems to have been settled with the kappa index of agreement as proposed by Cohen (1960; also see Swaminathan, Hambleton & Algina, 1974). In a recent work, Huynh (1977) argues for the assumption of test exchangeability (de Finetti, 1937; Hewitt & Savage, 1955; Zimmerman, 1975), and subsequently defines the population kappa based on this assumption. In another work, Huynh (1976) details the computation of kappa on the basis of data collected from one test administration when equivalent test data follow a bivariate beta-binomial model.

Kappa has been found to change with the cutoff score(s), hence there is no unique kappa for a given test administered to a group of examinees. Under test exchangeability, kappa varies between 0 and 1 inclusive. Hence there exists at least one set of cutoff score(s) at which kappa is <u>maximized</u>. We propose to call the upper bound of kappa the <u>kappamax</u> (KM) reliability index for the test (and for the group of examinees) under consideration.

This study will deal mainly with kappamax based on binary classifications. The beta binomial model will be used extensively to describe the test score distribution. Such model would assume that the test is composed by randomly drawing items from an item universe featuring the instructional objectives under study. This type of test is traditionally described as <u>domain-referenced</u>.

EVALUATION OF KAPPAMAX BASED ON THE BETA-BINOMIAL MODEL

Several computational procedures for the kappa reliability index at a given cutoff score are described at length in Huynh (1976). To obtain the kappamax, one may compute kappa for every cutoff score and then identify the maximum value that kappa can reach. (It is assumed, of course, that only nonrandomized classifications are to be used.) As an illustration, consider the situation where a five-item test is administered to a group of examinees with true ability defined by the beta density with parameters $\alpha = 11.43$ and $\beta = 6.80$. The values of kappa at the cutoff scores of 1, 2, 3, 4 and 5 are respectively .034, .095, .144, .148 and .091. Hence the value of kappamax is KM = .148.

The scheme described above is a sure but lengthy way to get at kappamax. A fairly efficient algorithm may be formulated by noting that (in all situations considered by the author up to now) kappa is an upturn U-shape function of the cutoff score. Moreover kappamax is usually reached at a cutoff score very near the test mean score. (See Table 1 for some typical numerical illustrations.)

Table 1

Thus the search for kappamax may be confined to cutoff scores in the neighborhood of the test mean score. The steps to be followed are:

(a) Choose two consecutive cutoff scores c and c+1 near the test mean score, and compute the corresponding kappa indices $\kappa(c)$ and $\kappa(c + 1)$.

(b) If $\kappa(c)$ is less than $\kappa(c+1)$, then compute $\kappa(c+2)$. If $\kappa(c+1)$ is larger than $\kappa(c+2)$, then the computation stops and the kappamax is

$\kappa(c+1)$. Otherwise, compue $\kappa(c+3)$, and so on.

(c) If $\kappa(c)$ is larger than $\kappa(c+1)$, then compute $\kappa(c-1)$. If $\kappa(c-1)$ is less than $\kappa(c)$, then the computation stops and the kappamax is $KM = \kappa(c)$. Otherwise, compute $\kappa(c-2)$, and so on.

AN APPROXIMATION FOR KAPPAMAX FOR LONG TESTS

The computations described in the last section for kappamax are rather tedious when the number n of test items is fairly large. Unless computer facilities are available, it seems desirable to seek an approximation for this case.

In a theoretical study regarding kappa based on normality, Huynh (1977) proves that the value of kappamax is

$$KM = (2/\pi)\sin^{-1}\rho \,, \tag{1}$$

where ρ is the traditional reliability. It follows that kappamax may be approximated by replacing ρ by a suitably chosen quantity representing the correlation between two equivalent forms of the tests. Several possibilities including the correlation $\rho = \alpha_{21}[(n-1)/(n+\alpha_{21})]^{\frac{1}{2}}$ (See Huynh, 1976, Formula (22)) have been tried, but it turns out that a reasonable approximation for kappamax may be obtained as $KM = (2/\pi)\sin^{-1}\alpha_{21}$, α_{21} being the traditional KR21 reliability index. It may be recalled that

$$\alpha_{21} = \frac{n}{n-1}\left[1 - \frac{\mu(n-\mu)}{n\sigma^2}\right]$$

where μ is the test mean score and σ^2 is the test variance. When sample data are to be used, α_{21} may be estimated by replacing μ and σ^2 by the corresponding sample mean $\bar{x}$ and sample variance s^2. To illustrate the approximation process, consider for example the

Duncan data referred to in Huynh (1976). Here $n = 20$, $\bar{x} = 12.54$ and $s = 3.05$. An estimate for the KR21 may be taken as $\hat{\alpha}_{21} = .52$, and hence KM ≃ .35. Exact calculations based on the beta-binomial model yield the value of .36 for kappamax.

Table 2 provides moredata on the approximation of kappamax based on normality. Three sets of true ability distributions are selected to represent different degree of homogeneity and four levels of test lengths are chosen to be n = 10, 20, 30 and 40. Set 1 yields unimodal and fairly homogeneous test score distributions. Set 2, on the other hand, may be taken as representative of unimodal distributions with moderate degree of homogeneity. Finally, Set 3 is a collection of bimodal and relatively heterogeneous test score distributions. For all the situations under consideration, the approximation based on normality provides kappamax slightly smaller than the corresponding true values. However, the absolute error (difference between a true kappamax and its approximate value) <u>never</u> exceeds six units in the second decimal. The mean absolute error stands at .023 at the relative error averages out at 3.7 percent. The data of Table 2 also reveals that the approximation becomes better with more test items. However, even with only ten items, the errors are still relatively small and perhaps negligible for most practical purposes.

Table 2

FACTORS AFFECTING KAPPAMAX

Since kappamax (based on binary classifications) does not involve the cutoff score, only the number n of test items and the true ability distribution (reflected by α and β) will have to be considered.

<u>Effect of test length</u>. Test reliability ρ is known to be an increasing function of n, hence kappamax defined under normality by $KM = (2/\pi)\sin^{-1}\rho$, is also an increasing function of n. This statement should be expected to hold asymptotically for the beta-binomial test score model. To shed light on the behavior of kappamax based on tests with short or moderate lengths, several computations are made and the results are compiled in Table 3. The data vividly document the increasing variation of kappamax with respect to n. Moreover they also indicate that the rate of increase becomes smaller as n takes larger values. The trend is reminiscent of the one induced by the Spearman-Brown formula. A later section of this paper will describe a way to project kappamax for tests of doubled length.

<u>Effect of true (or test) score variability</u>. The relationship between the true (or test) score variability and kappamax may be inferred from the data of Table 3. It may be noted that for the first two classes of test score distributions, kappamax varies in the same direction with the standard deviation σ_θ of the true score (or ability) distribution. This observation holds for each of the test length levels. As for the third type of test score distributions, namely those with two modes located at both extreme ends of the score range, the relationship between kappamax and σ_θ is not totally straighforward. It may be stated, however, that as a rule, kappamax tends to increase along with the degree of heterogeneity in the true (or test) score distribution.

Table 3

SAMPLING DISTRIBUTION OF KAPPAMAX

So far we have considered mostly the case where both parameters α and β of the beta distribution are known accurately. In most practical situations, however, these quantities are to be estimated from sample data. There are many way to accomplish this task. Moment estimates for α and β may be obtained by replacing the test mean μ and the test standard deviation σ by their corresponding sample statistics $\bar{x}$ and s in the formulas defining α and β. Thus the moment estimates are expressed as

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\,\bar{x},$$

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n,$$

where

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left[1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right].$$

These quantities are fairly simple to compute. However, in sampling studies, it is usually more desirable to consider the maximum likelihood (ML) estimates for their statistical inference properties are well known at least asymptotically (e.g. when the sample has a reasonably large number of cases.) It may be noted that for large samples, the moment estimates and their corresponding ML counterparts are very near each other in most cases.

An asymptotic statistical inference theory for kappa and kappamax is provided in Huynh (1977b) for beta-binomial test score distributions. To be specific, let $\widetilde{KM}$ be the ML estimate for the population kappamax KM. Then $\widetilde{KM}$ may be obtained simply by replacing the two parameters α and β in the formula defining kappamax by their ML estimates previously mentioned. If the estimation is based on N subjects, then

$N^{1/2}(\widetilde{KM} - KM)$ has an approximate normal distribution with zero mean and standard deviation σ_{KM}. Standard methods for evaluating this constant may be found in such text as Rao (1973) for example. Charts developed by Huynh (1977b) may also be used for this purpose. It follows from this discussion that the ML estimate $\widetilde{KM}$ is asymptotically unbiased for the population value KM and has a standard error equal to $\sigma_{KM}/N^{1/2}$.

The availability of a standard error for $\widetilde{KM}$ would serve at least two purposes. First, it may be used to establish confidence intervals for kappamax. For example, consider the case where N = 100, $\hat{\alpha} = 3.49$ and $\hat{\beta} = 1.08$. Recall that the moment estimates and their ML counterparts are very close to each other for large N's. Numerical computations provide the estimated kappamax of .395 with an estimated standard error of $.699/(100)^{1/2} = .070$. An approximate 95% confidence interval for the population kappamax may now be taken as $.395 \pm 1.645 \times .070$, which is the interval from .28 to .51. Second, the value of σ_{KM} may serve as a starting point for decisions regarding the number of subjects to be included in any kappamax estimation. For example, a pilot administration of a 10-item test to a group of examinees yields the values $\hat{\alpha} = 5$ and $\hat{\beta} = 3$. It follows that an estimate for σ_{KM} is .539. If a standard error of .02 is acceptable, then the number N of subjects should be at least $(.539/.02)^2$ or 726 approximately.

Table 4 presents some data regarding the variation trend of σ_{KM} with respect to α, β and n. It is clear that for a given group of examinees (e.g. for a given set of α and β), σ_{KM} is a decreasing function of n. It follows from previous discussions that σ_{KM} is also a decreasing function of KM. Across different groups of examinees, however, σ_{KM} and KM relate to each other in a rather unpredictable manner.

Table 4

PROJECTION FOR KAPPAMAX OF TESTS WITH DOUBLED LENGTHS BASED ON BETA-BINOMIAL DATA

Huynh (1977) shows that under normality, kappa is asymptotically a function of $n^{1/2}$, n being the number of test items. This suggests that an adaptation of the Spearman-Brown formula might be stated as

$$KM_1 = \frac{h^{1/2}\, km}{1 + (h^{1/2} - 1)\, km} \qquad (2)$$

where km is the kappamax index for a short test and KM_1 is the kappamax index projected for a test whose length is h times longer than the short test. If the original Spearman-Brown formula were to be used, then the predicted kappamax would be

$$KM_2 = \frac{h\, km}{1 + (h - 1)\, km} . \qquad (3)$$

It turnsout surprisinglythat for the 45 beta-binomial situations considered as representatives of a wide range of testing conditions, Formula (2) always underprojects kappamax whereas Formula (3) always overprojects this index for tests with doubled lengths (h = 2). This suggests that the compromised formula

$$KM = \frac{h^{3/4}\, km}{1 + (h^{3/4} - 1)\, km} \qquad (4)$$

would do a better projection. It fact it <u>does</u>. In all the 45 cases previously mentioned, the difference between the true value of kappamax and its projected value obtained from Equation (4) (e.g. the absolute error) <u>nevers</u>, exceeds two units of the second decimal. Overall the mean absolute error is .013 and the relative error (e.g. the ratio of the absolute error to the true kappamax) averages out at 1.96 percent. As may be seen from Table 5, the errors do not seem to relate to the level of kappamax from which projection is to be made.

Table 5

PROJECTION FOR KAPPAMAX OF TESTS WITH DOUBLED LENGTHS BASED ON REAL DATA

The beta-binomial model has been used in this study mainly because of its mathematical simplicity. Real test data, however, rarely conforms exactly to any parametric form. It seems therefore desirable to investigate the accuracy of the projection formula (4) using real test data.

The results subsequently presented in this section are based on the responses of 582 examinees to a 138-item test. The test items have difficulty indices ranging from .259 to .861 with a mean of .556 and a standard deviation of .142. Three levels of test length, namely n = 10, 20 and 40, are considered and thirty independent projections are made for each n. The following steps are taken in each projection.

(a) Four subtests, each with n/2 items are <u>randomly</u> selected from the 138-item pool, and four subscores namely X_1, X_2, Y_1 and Y_2 are computed for each of the 582 subjects.

(b) The sums $X = X_1 + X_2$ and $Y = Y_1 + Y_2$ are computed and treated as scores obtained from two equivalent forms of an n-item test. The true kappamax is then computed using the (X,Y) bivariate distribution and the average marginal distribution of X and Y.

(c) Two "split-half" kappamax, one based on (X_1, X_2) and the other on (Y_1, Y_2) are computed as in the previous step, and then projected via Formula (4) using h = 2. The average of the two projected kappamax is finally considered as the projected kappamax for the n-item test described in the previous step.

Table 6 reports the summary data compiled from the 90 projections. In this table, the quantity L_{mvc} (Gulliksen, 1950, page 175) expresses the extent to which the scores X_1, X_2, Y_1 and Y_2 have the same mean, the same variance, and the same covariance. When this similarity is fulfilled, L_{mvc} is unity. A small value for this quantity, on the other hand, indicates that the four subtests represented by X_1, X_2, Y_1 and Y_2 depart appreciably from equivalence. Projections based on the adapted Spearman-Brown formula would therefore not be expected to provide good results when L_{mvc} is too far from unity.

The data of Table 6 clearly show that Formula (4) leads to reasonably good projections for kappamax of tests with doubled lengths. Over the 90 cases under study, the absolute errors average out at .040 and the relative errors do at 6.47%. As expected, very good projections are obtained (with absolute errors not more than .05) when the split-half procedure provides reasonably equivalent subtests. As elaborated in Huynh (1976), a high degree of subtest equivalence may be secured by pairing the items by difficulty before randomly splitting each pair in the subtests.

Table 6

SOME CONCLUDING REMARKS

Several aspects regarding the kappamax reliability index for binary classifications are considered in this study. The advantage of the use of kappamax instead of an ordinary kappa lies in the fact that kappamax is not a function of the cutoff score. Hence kappamax is unique for a test given to a group of examinees. In some sense, kappamax reflects the situation in which a test functions best in terms of consistency of decisions. It should be mentioned that many aspects of statistical decision theory (such as the minimax principle) are actually based on the best performance of a procedure over a wide range of conditions. The asymptotic error prescribed for kappamax would be helpful in deciding the number of examinees to be included in a kappamax reliability study. The adapted Spearman-Brown procedure as reflected by Formula (4) would be useful in approximating the kappamax directly from the test data without imposing a parametric form for the test score distribution. We would like to caution the reader that the projection might not be appropriate if it is based on a small number of subjects. As indicated by Huynh (1976), negative values for various kappa might be a result in this case and their interpretation is rather messy. Finally there is no reason that kappamax must be defined as the maximum kappa over the whole range of test scores. If there are good reasons to restrict the range of possible cutoff scores, the appropriate "kappamax" may then be defined. It is expected that most of the properties mentioned for kappamax as defined in this study would hold for other "kappamax" based on a limited range of cutoff scores.

REFERENCES

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

De FINETTI, B. La prevision: ses lois logiques, ses resources subjectives. Annales de l' Institut Henri Poincare, 1937, 7, 1-68.

GULLIKSEN, H. Theory of mental test scores. New York: John Wiley & Son, 1950.

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.

HEWITT, E & SAVAGE, L. J. Symmetric measures on Cartesian products. Transactions of the American Mathematical Society, 1950, 80, 470-501.

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

HUYNH, H. Reliability of multiple classifications. Paper presented at the Spring meeting of the Psychometric Society. Murray Hill, New Jersey, April 1976. Also resubmitted to Psychometrika at the suggestion of the managing editor, March 1977.

HUYNH,H. Statistical inference for the kappa and kappamax reliability indices based on the beta-binomial model. Paper to be presented at the annual meeting of the Psychometric Society. University of North Carolina, June 16-17, 1977.

RAO, C. R. Linear statistical inference and its applications. New York: John Wiley & Sons, 1973.

SWAMINATHAN, H., HAMBLETON, R. K. & ALGINA, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 262-267.

ZIMMERMAN, D. W. Probability spaces, Hilbert spaces, and the axioms of test theory. Psychometrika, 1975, 40, 395-412.

Table 1

Variation of Kappa as Function of Cutoff Score c for n = 10

α	β	μ	c = 1	2	3	4	5	6	7	8	9	10
11.43	6.80	6.27	.012	.045	.095	.153	.204	.236	.242	.215	.154	.071
3.09	.98	7.59	.151	.269	.358	.426	.476	.512	.533	.538	.518	.437
1.22	.98	5.54	.423	.544	.602	.633	.649	.654	.648	.627	.582	.476
.30	.30	5.00	.749	.809	.832	.843	.847	.847	.843	.832	.809	.749

Table 2

A Comparison Between Exact Kappamax And Its Approximate Value Based on Normality

α	β	n	α_{21}	Exact KM	Appr. KM	Error	Percent Error
11.43	6.80	10	.354	.242	.230	.012	4.96%
		20	.523	.361	.350	.011	3.05
		30	.622	.436	.427	.009	2.06
		40	.687	.490	.482	.008	1.63
3.49	1.08	10	.686	.515	.481	.034	6.60
		20	.814	.630	.605	.015	2.38
		30	.868	.689	.669	.020	2.90
		40	.897	.727	.709	.018	2.48
.80	.50	10	.885	.744	.692	.052	6.99
		20	.939	.814	.776	.038	4.67
		30	.959	.847	.817	.030	3.54
		40	.968	.866	.839	.027	3.12

Table 3

Variation of Kappamax as Function of Test Length and Test Score Variability

Description	α	β	σ_θ	n = 5	10	20	40
Unimodal with mode not at the extreme scores	3.49	1.08	.180	.395	.515	.630	.727
	5.00	3.00	.161	.273	.395	.523	.639
	11.43	6.80	.110	.148	.242	.361	.490
	10.00	10.00	.109	.142	.224	.341	.471
Unimodal with mode at the extreme score	1.00	1.00	.289	.567	.670	.758	.825
	1.22	.98	.278	.546	.654	.745	.815
	3.09	.98	.190	.420	.538	.650	.742
	5.00	1.00	.141	.336	.460	.582	.688
Bimodal with modes at the extreme scores	.30	.30	.395	.791	.847	.891	.922
	.60	.50	.344	.688	.769	.833	.880
	.80	.50	.287	.654	.744	.814	.866
	.80	.20	.200	.710	.785	.844	.889

Table 4

Variation trend of σ_{KM} with Respect to n

α	β	σ_θ	n =10	20	30	40
11.43	6.80	.18	.692	.531	.450	.399
3.49	1.08	.11	.468	.328	.267	.226
.80	.50	.29	.269	.176	.139	.119

Table 5

Projection Errors for Kappamax of Tests With Doubled Lengths Based on Beta-Binomial Data

Range of km	No of Proj.	Mean of Rel. Error	Mean of Abs. Error
.000 - .200	3	6.53%	.006
.201 - .400	7	1.69	.007
.401 - .600	8	1.74	.011
.601 - .800	18	1.91	.015
.801 - .999	9	1.34	.012

Table 6

Projection Errors for Kappamax of Tests With Doubled Lenghts Based On Real Data

n	Range of L_{mvc}	No of Proj.	Mean Rel. Error	Mean Abs. Error
10	.59 or lower	4	22.65%	.116
	.60 - .69	3	18.19	.161
	.70 - .79	8	9.06	.048
	.80 - .89	11	2.23	.012
	.90 or higher	4	5.46	.030
		Overall	8.88	.047
20	.49 or lower	3	14.82	.080
	.50 - .59	7	7.22	.045
	.60 - .69	7	5.63	.035
	.70 - .79	11	4.13	.028
	.80 or higher	2	1.62	.011
		Overall	6.10	.038
40	.39 or lower	3	12.89	.096
	.40 - .49	2	4.44	.035
	.50 - .59	4	4.57	.036
	.60 - .69	6	6.10	.044
	.70 or higher	15	2.04	.018
		Overall	4.43	.035