Hagen, Randi Louise

AN EXPLORATION OF DECISION CONSISTENCY INDICES FOR ONE FORM TESTS

Iowa State University PH.D. 1983

An exploration of decision consistency indices for one form tests

by

Randi Louise Hagen

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Psychology

Approved:

Signature was redacted for privacy.
In Charge of Major Work

Signature was redacted for privacy.
For the Major Department

Signature was redacted for privacy.
For the Graduate College

Iowa State University
Ames, Iowa

1983

TABLE OF CONTENTS

	Page
DECISION THEORY AND EDUCATIONAL TESTING	1
CRITERION-REFERENCED TESTS	7
Psychometric Issues	9
Current Focus	13
CONSISTENCY INDICES: RHO AND KAPPA	15
Two Administration Indices	16
One Administration Indices	23
Empirical Comparisons	33
PURPOSE	38
STUDY I: SIMULATION STUDY	39
Data Generation	39
Distributions	39
Examinees	41
Test Lengths	41
Standards	42
Rho and Kappa Estimates	43
STUDY II: ACTUAL PLAN	44
Data Collection	45
Subtests	47
Examinees	47
Distributions	48
Standards	49

	Page
Rho and Kappa Estimates	50
RESULTS	51
Study I: Simulated Data	51
Study II: Actual Data	67
Comparisons of Studies I and II	82
DISCUSSION	86
Current and Previous Research	88
Implications for Practitioners	90
Implications for Future Research	95
REFERENCES	99
ACKNOWLEDGMENTS	105
APPENDIX A: BETA-BINOMIAL ASSUMPTIONS	106
APPENDIX B: GOODNESS-OF-FIT TEST FOR THE BETA-BINOMIAL MODEL	109

DECISION THEORY AND EDUCATIONAL TESTING

Measurement tools are widely used for the assessment and monitoring of learning and teaching. The value of concise, if not mathematical, means of communicating information is apparent. Being able to specify the achievement level of a student ("He is an A student"), even more precise information ("She answered 20 of 25 items correctly"), or information about the measurement tool itself ("Ten problems involving addition of two digits") is important in the ongoing process of student learning.

It is hardly arguable that a measurement instrument must be well-standardized and yield information that is valid, reliable, and useful. The values of standardization, though often overlooked, are many. Standardization provides objective, independent and repeatedly verifiable information. In addition, it offers detailed, quantifiable information, which can be subjected to mathematical scrutiny and analyses. Communication and economic benefits are also apparent, information can be passed along to others with a common agreement about its interpretation, and, once instruments are fully developed and standardized, useful information can often be collected and used with minimal expenditure of time and money (Nunnally, 1978).

Testing in education takes many forms, some well-

standardized and others less so: the Friday afternoon spelling bee in third grade, vocational interest inventories in the guidance office, annual achievement test batteries, the senior high final examination. The role that testing plays in education also varies. It can be descriptive in nature, or designed to assess specific goals or objectives (Brown, 1983). However, the common element in all educational testing is its role in decision-making processes.

Decision-making in education occurs in daily instruction (e.g., composition of reading groups, choosing today's instructional modes or curricula for a given student), as well as in more formal contexts such as promotion or retention policies or ability grouping. While teacher-made tests play a major role in daily classroom decisions, standardized tests are an increasingly common facet of educational policies affecting large numbers of students.

Decision theory offers a paradigm for looking at educational testing in this light. Figure 1 provides a graphic representation of the decision-making process.

"Information", in the above figure, refers to any data which is used in decision-making, while a "strategy" is a formalized rule for using the information to arrive at a "decision" or course of action.

This schema is easily translated into an educational

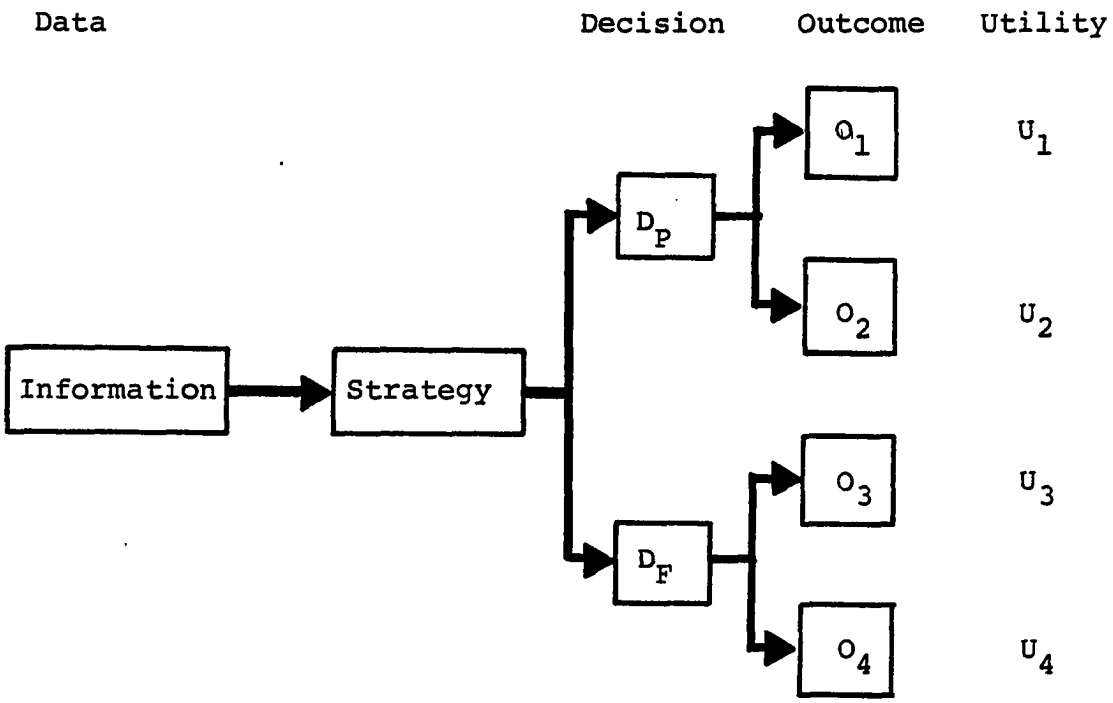

Figure 1. Representation of the decision-making process

testing context. Test scores provide the information component, either in concert with other information or alone. The strategy refers to a rule applied to the information, e.g., a grading scheme ("90%+ for an A, 80%+ for B". . .), a standard ("70 points is passing"), or an interpretation scheme ("a score of 10 on scale L indicates high math anxiety"). A decision can be based on the application of the strategy to the information, for example, "Jan earned 75 points; the passing standard was 70 points; Jan passed the

test and can, thereby, move on to the next level".

Once a decision is reached, an "outcome" occurs. For example, in the situation where Jan earned a passing score, let us presume that the test taken was designed to determine admission to an advanced class in algebra. Two decisions were possible: (D_P) Jan passed or (D_F), Jan failed. Two possible states of nature are also possible: Jan does or does not perform adequately in the class. Four combinations of decisions and states are, thus, possible: Jan passes the test and does well in class (O_1); Jan passes the test and fails in class (O_2); Jan does not pass the test but would have done well in class (O_3); Jan does not pass the test but would have failed the class (O_4).

In most contexts, not all outcomes are equally desirable. Certain combinations of decisions and states of nature are preferable to others; utility functions specify the relative degree of desirability for each outcome. In our example, Jan passing and performing adequately is most likely the preferred outcome, both by Jan and the school. Utility functions specify the degree of acceptability or preference for all four outcomes, including "false positives" (O_2) and "false negatives" (O_3). The goal of the process is, clearly, to maximize preferable outcomes; we want to reach the "best" decision regarding Jan's admission to the algebra class. More generally, the goal in any decision process is the

maximizing of utility functions by reaching the "right" decisions.

The use of tests in making "best" decisions in education has greatly expanded in recent years. Educational tests and their uses have received increased attention, as issues of test validity, accountability, and centralization of control have been in the spotlight (Haney, 1981). Controversies involving the role of tests in decisions regarding promotion and retention (e.g., Beckham, 1980), placement of students in special education (e.g., Reschly, 1981), and minimum competency testing (e.g., Resnick, 1980) are prevalent.

With this changing and increasing role of testing in educational decision-making, technical concerns have also surfaced. Tests used for decision-making are increasingly those designed for that specific purpose and the use of criterion-referenced tests has greatly increased. Theoretical and psychometric progress has had to keep pace with the changing use of tests, and, indeed, one can cite volumes of research on such issues (e.g., Berk, 1980; Shepard, 1980).

Wiggins (1973) has cogently and emphatically stated a major philosophical and psychometric concern of criterion-referenced tests:

> From a practical standpoint, the number of correct decisions made by a . . . test or assessment, is a more important piece of information than the degree of association that exists between predicted and obtained scores (p. 230).

In this light, the current investigation focuses on the issue of consistency of decisions. Given specific information and a strategy, how can we index the degree of consistency with which we reach a decision? The concern is with indices estimating the degree of consistent classification of students from scores on criterion-referenced tests.

CRITERION-REFERENCED TESTS

Criterion-referenced tests are not a new phenomenon in education. In 1864, Rev. G. Fisher developed a proficiency scale (1=best, 5=poorest) for academic performance in writing, spelling, grammar, composition and mathematics (Chadwick, 1864, in DuBois, 1970). A "graphometer" was published by E. L. Thorndike in 1910 to measure handwriting with equal unit scaling and a level deemed to be minimum proficiency (DuBois, 1970).

Tests with a criterion of proficiency have long been used in the classroom - a teacher decides that a score of 65 is needed to pass an arithmetic test, or students must spell 8 out of the 10 new words correctly to be awarded the Good Speller of the Week award. Use of tests with set standards for passing or failing, or specified degrees of proficiency, have been and are a common occurrence in American education.

Glaser and Klaus (1962) first used the term criterion-referenced test in connection with tests setting standards of proficiency in industrial training. The following year, Glaser (1963) expanded the concept:

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge aquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on the continuum as indicated by the

> behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. . . . Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others (p. 519).

In the two decades since this seminal article, criterion-referenced tests have come out of the classroom and into the spotlight in educational testing. Hundreds of references to criterion-referenced tests are seen in the literature (see, for example, Hambleton, Swaminathan, Algina & Coulson, 1978). The popularity of objectives-based education (e.g., Popham & Baker, 1970); Mager, 1972), mastery learning (e.g., Bloom, 1971), and the minimum competency movement (see Resnick, 1980; Lerner, 1981) have been factors in the increased use of criterion-referenced tests.

Although there is no one prototypical criterion-referenced test (Nitko, 1980), a criterion-referenced test is distinguishable from other tests in that it is "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971, p. 653). This is re-iterated in Popham's (1975) statement that "a criterion-referenced test is used to ascertain an individual's status (referred to as a domain score) with respect to a well-defined

behavior domain" (p. 130). A large literature has been generated with respect to the definition and delineation of the term "criterion-referenced" (see, for example, Millman, 1974, or Hambleton & Novick, 1973). The term used herein is in keeping with the above cited definitions of Glaser & Nitko (1971) and Popham (1975).

Psychometric Issues

The psychometric issues of validity and reliability must be taken into account in judging the worth and value of a criterion-referenced test as well as other types of tests (Standards for educational and psychological tests, 1974).

Validity

Berk (1980) discusses the concerns regarding validity of criterion-referenced tests: content validity, the validity of scores for intended use, and validity of classification. The first, content validity, refers to the match between the test content and the objectives of the test, i.e., does the test contain items which measure the objective(s) it intended to measure? The second concern refers to questions regarding use or interpretation of scores for their intended use. Messick (1975) and Linn (1979) argue cogently for the construct validity of criterion-referenced tests, for the necessity for evidence about their

proper interpretation and use.

Berk's last concern, validity of classification, involves the match between test scores and classification of examinees based on their scores, i.e., is the standard set at a score which truly distinguishes between classifications of examinees? A large and controversial literature surrounding the standard setting question has developed (e.g., Zieky & Livingston, 1977; Glass, 1978; Skakun & Kling, 1980) in this regard.

Reliability

Reliability is a generic term referring to the consistency of performance over samples of items and testing occasions (Brown, 1980). The classical definition of reliability is "the measure of the degree of the true-score variation relative to the observed-score variation" (Lord & Novick, 1968, p. 61). Coefficients of reliability have been developed to indicate this ratio of true to observed variance according to the desired or prescribed comparison (e.g., coefficient alpha for inter-item comparisons, coefficients of stability for time comparisons, or coefficients of equivalence for comparing forms); the commonality of coefficients lies in their indication of the degree of consistency of scores, whether across items, time or forms.

The calculation of the reliability of a test, in the classical sense, is based on the variability of scores. This

is true for criterion-referenced tests as well as others, and a single coefficient or index of consistency can be calculated. Hambleton, Swaminathan, Algina & Coulson (1978) discuss the various methods for estimation of a true score or, in the case of criterion-referenced tests, domain score. An obvious concern is simply the estimation of a domain score without regard to a standard(s). The standard error of measurement is a commonly found index in this regard, and is applicable to criterion-referenced tests as well as other types of tests.

Popham & Husek (1969) commented that since criterion-referenced tests are often used in situations where the instructional intention is to maximize the number of students in the mastery category, thus, minimizing variance, the use of the traditional concept of reliability - the ratio of true to observed score variance - in inappropriate.

Criterion-referenced tests pose another unique consideration in terms of the classical concept of reliability: the major concern for consistency often lies not with the consistency of an individual's score itself, but with the consistency of classification of the individual with respect to a standard. To be specific, the question is "Is student X in the same mastery category on both forms/ administrations?" rather than "Does student X have the same score on both forms/administrations?" Although this form of

consistency clearly has a place under the rubric reliability, this distinction from the classical "ratio of true to observed variance" definition is important. The terms, reliability, agreement, and consistency index have all been used in this regard (Berk, 1980). For the sake of clarity, the term "consistency" shall be hereafter used when referring to agreement of classification rather than agreement between test scores.

Hambleton, Swaminathan, Algina & Coulson (1978) delineated three major categories of reliability to be considered with criterion-referenced tests. One, within the classical framework, is the estimation of the domain (true) score. The authors refer to the other two types of reliability as "reliability of criterion-referenced test scores" and "reliability of mastery classification decisions". Both are concepts of consistency in which the relationships between scores and standards are crucial. The topic to be explored herein is the latter, consistency of mastery classification decisions. The prior concept, reliability of criterion-referenced test scores, refers to the consistency of squared deviations of individual scores from the standard and is analogous to deviations from the mean.

The major distinction between the two standard-related concepts of consistency is the judged seriousness of classification errors for individual test-takers. For example,

suppose you have two forms of a 10 item test with a mastery standard of 8, and that student 1 scores 8 on form A and 7 on form B while student 2 scores 9 on form A and 4 on form B. Both students were, thus, masters on form A and nonmasters on form B. The discrepancy for student 1, however, was between scoring at standard versus 1 point below the standard, while for student 2 the discrepancy was 1 point above standard versus 4 points below. In the first sense of consistency, the size of the discrepancy between scores is a factor: the student with the greater discrepancy is regarded as a more serious inconsistency in classification. Indices which measure the latter concept of consistency of criterion-referenced test scores (squared error loss) have been developed by Brennan and Kane (1977) and Livingston (1972). In the latter concept of consistency, that of mastery classification, both inconsistencies are judged to be of equal seriousness: the degree of discrepancy between scores is irrelevant. What is of concern is classification consistency alone.

Current Focus

The concept of mastery decision consistency is the focus of the current discussion and research. This concept follows from the premise that all inconsistencies of classifica-

tion are of equal seriousness, whether resulting from a discrepancy of one or many points. Two coefficients have been developed as indicators of classification consistency, rho ($\hat{\rho}$) and kappa ($\hat{\kappa}$). The development and current status of these coefficients and two estimation procedures for use with tests with only one form are discussed below.

CONSISTENCY INDICES: RHO AND KAPPA

Carver (1970) proposed two procedures for indicating the consistency of decisions regarding mastery classification: (1) a comparison of the percentage of students classified as masters/nonmasters on two parallel tests, and (2) a comparison of the percentage of masters/nonmasters on the same test administered to two matched groups. While providing an overall index of general consistency, neither of the procedures was sensitive to individual consistency. For example, 50% of the test-takers may be classified as masters on each form, but it is possible that every master on the first form was a nonmaster on the second form. In the second procedure, comparability of the matched groups is questionable; while testing of the groups may result in 50% masters in each group, nothing is known regarding the comparability of masters and nonmasters across the two groups.

Research since Carver's initial conceptualizations has focused the consistency of individual's classifications rather than merely the percentage of group masters and nonmasters (Berk, 1980.

Two Administration Indices

Two coefficients that take individual consistency into account are the rho ($\hat{\beta}$) coefficient adapted for this use by Hambleton & Novick (1973) and Swaminathan, Hambleton & Algina (1974), and the kappa ($\hat{\kappa}$) coefficient first developed by Cohen (1960) and adapted by Swaminathan, Hambleton and Algina (1974). Rho refers to the proportion of individuals consistently classified as masters or nonmasters on two parallel tests, while kappa refers to the proportion of consistent classifications beyond the chance level.

Table 1 displays data for 30 students on parallel forms of a ten item test (Subkoviak, 1980). Although the number of items and examinees are small relative to the type of tests discussed herein, this data set will serve as an example throughout this chapter. It is also noted that the scores in the example show a large proportion of nonmasters, and most criterion-referenced tests are used in anticipation of a large proportion of masters. The formulae and calculation of the consistency indices are, however, not affected by this skewness.

Rho indicates the proportion of individuals consistently classified as masters or nonmasters. As can be seen in Table 1, with a mastery level of 8 correct, student 2 was the only master on both forms A and B, and students 3 through 30 were

Table 1. Scores of 30 students on two forms of a ten-item test (Subkoviak, 1980)[a]

Student	Form A	Form B
1	9	7
2	8	8
3	7	7
4	7	4
5	7	3
6	6	7
7	6	7
8	6	5
9	6	4
10	5	6
11	5	4
12	5	2
13	5	2
14	4	7
15	4	7
16	4	7
17	4	6
18	4	4
19	4	4
20	4	4
21	4	3
22	4	2
23	3	6
24	3	4
25	3	4
26	3	4
27	3	2
28	3	2
29	2	4
30	1	1

[a]Mastery level = 8 correct.

consistently nonmasters. Student 1 was a master on form A and a nonmaster of form B. A tabular display of the consistency of classification can be seen in Table 2. A total of 29 of the 30 students (the diagonal cells) were consistently classified on both forms. To calculate the rho

Table 2. Mastery and nonmastery consistency for scores in Table 1

Form A	Form B		TOTAL
	Mastery	Nonmastery	
Mastery	1	1	
Nonmastery	0	28	
TOTAL	1	29	

coefficient, the proportion of individuals consistently classified, the formula below is

$$\hat{\rho}_o = \sum_{k=1}^{m} \hat{p}_{kk}$$

where $\hat{p}_{kk}$ = proportion of individuals consistently classified in the k^{th} mastery category on both tests, and m = number of categories.

For the data in Table 1, the proportion of individuals consistently classified is:

$$\hat{\rho}_o = \frac{1}{30} + \frac{28}{30} = \frac{29}{30} = .97.$$

If all individuals are consistently classified, rho reaches an upper limit of +1.00. The lower bounds of rho are determined by the chance level; except for highly unusual circumstances, the lowest rho coefficient would be that seen if classification as master or nonmaster was purely random. The chance level is:

$$\hat{\rho}_c = \sum_{k=1}^{m} \hat{\rho}_{k\cdot} \hat{\rho}_{\cdot k},$$

where $\hat{\rho}_{k\cdot}$ and $\hat{\rho}_{\cdot k}$ are the proportion of individuals assigned to the respective mastery and nonmastery classification on each test form.

For the scores in Table 1:

$$p_c = (\frac{2}{30} \times \frac{1}{30}) + (\frac{28}{30} \times \frac{29}{30})$$

$$= \frac{2}{900} + \frac{812}{900}$$

$$= .90.$$

Swaminathan, Hambleton and Algina (1974) suggested that the chance factor should be omitted from the index of consistency, as the index of interest is the consistency of individual classification due to the test alone. They suggested the use of Cohen's (1960) kappa:

$$\hat{\kappa} \quad \frac{\hat{p}_o - \hat{p}_c}{1 - \hat{p}_c}$$

where $\hat{p}_o$ = rho and $\hat{p}_c$ = chance level.

For the Table 1 data, we calculate:

$$\hat{\kappa} = \frac{.97 - .90}{1 - .90} = .70$$

These estimates, based on two actual test administrations, are hereafter referred to as two form estimates. As can be seen, the rho and kappa coefficients measure two different aspects of consistency, rho referring to consistent classification for any reason, and kappa to consistent classification beyond chance.

Comparing rho and kappa

Before discussing the development of estimates of rho and kappa based on one rather than two test administrations, it is valuable to gain a perspective on the similarities and differences between the two coefficients. As mentioned above, in both cases the upper limit is +1.00, which occurs when there is perfect agreement in classification, i.e., when every individual who is a form A master is also a form B master, and every nonmaster is such on both forms. In such a condition, Table 2 would have 0's in the off-diagonal cells.

The lower limits of the two coefficients, however, differ. The lower limit of rho is generally the proportion of consistent classification expected by chance. The lower limit of kappa, in contrast, is -1.00, a condition which would occur with perfect inconsistency, i.e., if all form A masters were form B nonmasters, and vice versa.

The role of marginals in the two coefficients differ as well. Although the degree to which a test is easy or hard (i.e., results in a large proportion of either masters or nonmasters) will determine the general degree of consistency, the marginals themselves are more crucial in the calculation of the kappa than the rho coefficient. The rho coefficient indicates the proportion of consistent classifications: if 24 of 30 students fall consistently in the same classification, whether the 24 consistent students are composed of 12 masters + 12 nonmasters or 22 masters + 2 nonmasters does not affect the rho coefficient. In both cases, rho is .80. Kappa, however, because it takes the proportion attributable to chance into account, a factor which is determined by marginals, will not be the same in the two above events. In the former (12 masters + 12 nonmasters), kappa is .50, and in the latter (22 masters + 2 nonmasters), kappa is .72. Furthermore, the marginals indicating inconsistent classification are also accounted for in the

calculation of kappa, but not rho. Kappa is, therefore, not a direct function of the value of rho.

This difference in rho and kappa, the role of the chance level, is a major factor in the decision of whether to use rho or kappa as an index of consistency. Livingston and Wingersky (1979) criticized the use of the kappa coefficient because of the role of the correction for chance:

> Applying such a correction to a pass/fail contingency table is equivalent to assuming that the proportion of examinees passing the test could not have been anything but what is happened to be. For example, if 87% of the examinees passed the test, kappa will "correct for chance" under the assumption that "chance" would result in exactly 87% of the examinees passing the test. This assumption makes sense when the pass/fail cutoff is chosen on the basis of the scores to which it will be applied, so as to pass a specified proportion of the examinees. It does not make sense when the pass/fail cutoff represents an absolute standard that is to be applied individually to each examinee (p. 250).

The choice of kappa or rho is a decision partially based on whether chance is to be included in the concept of consistency of mastery classification; does one want to know the consistency of classification regardless of the source of that consistency or does one want to know the consistency attributable to the test alone? The technical or psychometric behavior of the coefficients, discussed below, must also be taken into account.

One Administration Indices

It is not uncommon that criterion-referenced tests have only one rather than two or more parallel forms. Huynh (1976) and Subkoviak (1976) responded to this concern in separate developments of estimates of rho and kappa coefficients which can be calculated from one test administration. In both cases, their methods involve the actual administration of the one available test form and the simulation of scores on a second (hypothetical) form. Hence, from the two sets of scores (one actual and one hypothetical), estimates of rho and kappa can be calculated as in the two-administration case discussed above.

The difference between Huynh's and Subkoviak's methods lies in the procedures for simulating the second form scores and the attendant assumptions. Both methods have gained increasing attention over the past few years, as a result of the increased use of and demand for psychometric information about criterion-referenced tests.

The Huynh (1976) and Subkoviak (1976) estimation procedures are not the only methods for calculating rho and kappa estimates with one test administration. Marshall and Haertel (1975) developed a procedure for simulating the second test scores based on the calculation of scores on a hypothetical double-length test (i.e., one with twice the number of items

on the administered form), splitting this hypothetical test into half-tests and calculating the consistency between the two halves. The mathematical complexity, lack of available computer programs and relative lack of research on the Marshall & Haertel procedures have deemed it the least applicable of the current estimation procedures, and it has not been included herein.

Huynh estimation procedure

Outlined below is Huynh's (1976) method for simulating a second administration, calculating rho and kappa coefficients, and a brief discussion of a second related approximation technique.

The gist of the Huynh estimation procedure is the simulation of a second hypothetical distribution of test scores by assuming a beta-binomial joint distribution between the actual and simulated distributions. That is, we can use the scores on form A, calculate the parameters (alpha and beta) of the form A (beta-binomial) distribution, and using these parameters, simulate form B scores. The key assumption is that of a beta distribution, which allows us to simulate form 2 scores based on scores of the one actual administration. Appendix A offers a brief description of beta distributions, alpha and beta parameters, and current research concerns regarding the beta-binomial model. The data used

in this and the following discussion of the Subkoviak procedure consist of the test scores on form A of Table 1.

The three steps for simulating form B scores are:

1. Sample statistics from form A scores: The mean ($\hat{\mu}$), variance ($\hat{\theta}^2$), and Kuder-Richardson 21 (KR_{21}) for scores on form A are 4.63, 3.27 and 0.27, respectively.

2. Distributional parameters: Parameters alpha ($\hat{\alpha}$) and beta ($\hat{\beta}$) are calculated from form A scores. These parameters reflect the first and second moments of the distribution and their significance lies in their determination (along with n) of the particular shape of the distribution of scores on the simulated form (see Appendix A).

$$\hat{\alpha} = (-1 + \frac{1}{KR_{21}})\hat{\mu} = (-1 + \frac{1}{0.27})4.63 = 12.52$$

$$\hat{\beta} = -\hat{\alpha} + \frac{n}{KR_{21}} - n = -12.52 + \frac{10}{0.27} - 10 = 14.52$$

3. Form B scores: Using the values of alpha, beta and the number of items, the joint distribution of scores can be determined. This two-step process involves the calculation of f(0,0), the probability of an individual scoring 0 on the simulated form given a score of 0 on form A, and the subsequent calculation of probabilities of all other combinations of scores. Computational formulae for the two steps are shown below:

$$\hat{f}_{(x,y)} = \prod_{i=1}^{2n} \frac{2n+\hat{\beta}-i}{2n+\hat{\alpha}+\hat{\beta}-i}$$

$$\hat{f}_{(x+1,y)} = f_{(x,y)} \frac{(n-x)(\hat{\alpha}+x+y)}{(x+1)(2n+\hat{\beta}-x-y-1)} .$$

The expected frequencies calculated are displayed in Table 3 below. Each entry is the proportion of examinees who would obtain score y on form B given a score of x on form A. For example, the frequency with which a score of 5 is expected on the simulated form B given a score of 3 on form A, is 0.0299. The table is symmetrical in that +(3.5) = f(5,3). Note that decimals have been omitted.

The simulation of form B scores allows for calculation of the consistency coefficients as if two actual forms had been administered. Rho can, thus, be obtained by summing the proportion of consistent classifications in Table 3. With a mastery level of 8, the proportion of individuals who are consistently classified as masters (lower right quadrant of matrix) is .0082. The proportion of individuals consistently classified as nonmasters is reached by summing all frequencies in the upper left quadrant (those scoring 7 and 7 and all combinations of lesser numbers), .8938. Rho is, thus, the total proportion of consistent masters and consistent nonmasters:

Table 3. Joint distribution of scores for forms A and B[a] (Subkoviak, 1980)

Form A scores	Form B scores										
	0	1	2	3	4	5	6	7	8	9	10
0	0002	0006	0011	0013	0012	0008	0004	0002	0000	0000	0000
1	0006	0024	0050	0069	0068	0059	0028	0012	0004	0001	0000
2	0011	0050	0116	0174	0188	0152	0093	0043	0014	0003	0000
3	0013	0069	0174	0286	0338	0299	0201	0101	0036	0008	0001
4	0012	0068	0188	0338	0436	0421	0308	0169	0066	0017	0002
5	0008	0050	0152	0299	0421	0444	0354	0211	0090	0025	0003
6	0004	0028	0093	0201	0308	0354	0308	0200	0093	0028	0004
7	0002	0012	0043	0101	0169	0211	0200	0142	0072	0024	0004
8	0000	0004	0014	0036	0066	0090	0093	0072	0040	0014	0003
9	0000	0001	0003	0008	0017	0025	0028	0024	0014	0006	0001
10	0000	0000	0000	0001	0002	0003	0004	0004	0003	0001	0000

[a]Each entry represents the proportion of examinees who would obtain score y on form B given score x on form A. Decimals have been omitted for ease in reading.

$$\hat{\rho}_o = .0082 + .8938 = .90$$

The kappa coefficient can be calculated in a manner similar to the two-form procedure discussed above. The proportion attributable to chance is a function of the marginal proportions based on Table 3. The proportion of masters on form A is the sum of the last three rows (.0577), and the proportion of nonmasters is 1 - .0057, or .9423. For our data:

$$\hat{\rho}_c = (.0577 \times .0577) + (.9423 \times .9423)$$

$$= .0031 + .8900$$

$$= .89.$$

Kappa, according to Equation 3, is:

$$\hat{\kappa} = \frac{.90-.89}{1-.89} = .09.$$

The two form rho and kappa coefficients (based on both administrations) were .97 and .70, respectively, while with one administration and utilizing the Huynh procedure, rho and kappa were .90 and .09, respectively. (The differences in the coefficients estimated will be discussed after the Subkoviak procedure has been outlined.)

Huynh offers a second approximation method which is computationally less complex (and thus less expensive and

tedious). Huynh's second approximation method involves the same steps as those shown above, with the addition of an arcsin transformation of scores in order to normalize the distribution. Computation of the normal deviate comparable to the standard, and calculation of the probabilities of scores less than this value are essential parts of the procedure. Peng and Subkoviak (1980) found that the elaborate arcsin transformation is not necessary and that a simple normalizing procedure is a better method. The addition assumption of normality of the joint distribution of scores renders this approximation method less robust and has not gained the research interest shown for Huynh's first procedure.

Subkoviak estimation procedure

Subkoviak (1976) developed a method for calculating the rho and kappa coefficients in a much less mathematically complex manner. Test scores for form A of Table 1 will remain the basis for the explanatory calculations. Table 4 depicts the estimation process described thereafter.

1. Columns 1 and 2 of Table 4 contain test scores and frequencies, respectively. The mean and KR-21, as previously calculated, are 4.63 and .27, respectively.

2. The assumption is made that the 10 item test is a

Table 4. Subkoviak estimation procedure with Table 1 data

1	2	3	4	5	6	7
X	N_x	$\hat{P}_x$	$\hat{P}_x$	$1-2(\hat{P}_x-\hat{P}_x^2)$	$N_x[1-2(\hat{P}_x\hat{P}_x^2)]$	$N_x\hat{P}_x$
9	1	.90	.930	.869	.869	.930
8	1	.80	.678	.563	.563	.678
7	3	.70	.383	.527	1.582	1.149
6	4	.60	.167	.721	2.887	.668
5	4	.50	.055	.896	3.584	.220
4	9	.40	.012	.976	8.786	.108
3	6	.30	.002	.996	5.976	.012
2	1	.20	.000	1.000	1.000	.000
1	1	.10	.000	1.000	1.000	.000
TOTAL	30				24.248	3.765

sample from the domain of all such items. Column 3 displays the estimates of the proportion of items in the domain that an individual which each test score is expected to answer correctly. This is the proportion of items correct on form A, the probability of a correct item response. (Subkoviak offers an alternative calculation of $\hat{\rho}$, a regression estimate. Though preferable with homogeneous groups of students, in the context of districtwide testing, the homogeneity assumption is unlikely to be met, and the proportion of correct items on form A are used.)

3. Column 4 indicates the probability of an individual's classification as a master. Test items are assumed to be trials in a binomial process and we wish to know the probability that in ten trials (items) an individual will make eight or more successes or items correct (see Tables of the Binomial Probability Distribution, 1949). $\hat{\rho}_x^2$ is the probability that an individual will be consistently classified as a master on 2 independent testings; the converse, that the student will be consistently classified as a nonmaster is $(1-\hat{\rho}_x)^2$. Column 5 shows the probability of consistent classification, the sum of the probabilities of the two classifications (master and nonmaster), $\hat{\rho}_x^2 + (1-\hat{\rho}_x)^2 = 1-2(\hat{\rho}_x-\hat{\rho}_x^2)$.

4. The probability of consistent classification across the entire group if displayed as the summation of column 6. Subkoviak's rho coefficient can be calculated by dividing this summation by n (30):

$$\hat{\rho}_o = \frac{\Sigma nx[1-2(\hat{\rho}_x-\hat{\rho}_x^2)]}{N}$$

$$= \frac{26.248}{30} = .91.$$

5. The summation of frequencies times the probability that an individual will be consistently classified on two independent testing is given by the total of column 7.

This is used in the calculation of the chance level:

$$\hat{\rho}_c = 1-2[\frac{\Sigma N_x P_x}{N} - (\frac{\Sigma N_x P_x}{N})^2]$$

$$= 1-2[\frac{3.765}{30} - (\frac{3.765}{30})^2] = .78.$$

6. With the chance level calculated, kappa can be easily obtained:

$$\hat{\kappa} = \frac{.89-.78}{1-.78} = .50.$$

The two-form, Huynh and Subkoviak estimates of rho and kappa are shown in Table 5. As can be seen, all three methods yield more similar estimates of rho coefficients than kappa coefficients.

Table 5. Comparison of rho and kappa estimates

	Rho	Kappa
Two forms	.97	.70
Huynh estimates	.90	.09
Subkoviak estimates	.89	.50

Empirical Comparisons

Although the development of rho and kappa and estimation procedures for one form administrations have generated much research (e.g., Huynh, 1979; Algina & Noe, 1978; Wilcox, 1981), only recently have studied addressing a comparison of the Huynh and Subkoviak estimates of consistency. Studies have compared the two estimation procedures for the rho coefficient with actual test data (Subkoviak, 1978), and rho and kappa were simulated test data (Marshall & Serlin, 1979).

Subkoviak (1978) compared the two-administration rho coefficient with the estimation procedures developed by Huynh (1976) and Subkoviak (1976). The study involved 1586 students, each of whom took parallel forms of a 50 item test developed from items on the Scholastic Aptitude Test (SAT). Ten and thirty item subtests were extracted and studied in addition to the fifty item test. On each of the three tests

(10, 30, and 50 items), four standards were considered: 50%, 60%, 70%, and 80% of items correct.

The percentages of students consistently classified on both forms of the 12 tests (3 lengths x 4 standards) were calculated and referred to as parameter values. Using both 50 classroom-size samples (n=30) and 50 larger samples (n=300), Subkoviak calculated rho coefficients using three methods: the two-administration method, the Huynh estimation procedure, and the Subkoviak estimation procedure. Comparisons of rho coefficients (parameter value versus mean estimates of rho calculated by each of the three methods) and standard errors were made for each of the 12 tests.

Overall, Subkoviak found standard errors (regardless of test length or placement of standard) of less than .08, although larger standard errors were seen with the two-administration method and the 10-item tests given to classroom size samples. The two-administration method (Swaminathan et al., 1974) produced, as expected, results in agreement with the parameter values.

A key finding of the study was the observed bias of the one-administration rho estimates, which appeared to be a function of test length and proximity of the standard to the mode. On the shorter tests, rho coefficients derived by the Huynh procedures were consistently lower than parameter values. Research by Huynh and Saunders (1979) also yielded

this underestimation of rho, as well as a similar bias with the kappa coefficient.

Coefficients calculated by the Subkoviak method showed a different pattern of bias: in short tests, underestimates were obtained when the standard was near the mode (50% of items correct), and overestimates when the standard was near the tails of the distributions (80% of items correct). Algina and Noe (1978) found a similar pattern of bias with the Subkoviak method.

Subkovick (1978) did not address the question of the shape(s) of the tests score distributions. Although it is implied that the distribution of all examinees' scores (i.e., all classroom samples, and all large samples) were normal, with data from one classroom sample provided as an example, it is doubtful if all samples resulted in normal distributions. Thus, the effect of distribution shapes on estimates of rho (and kappa) remain in question.

Marshall and Serlin (1979) did address the effect of the score distribution shape on rho and kappa estimates using simulated test data. Five distributions of 5, 10 and 20 items each were simulated: normal, left-skewed unimodal, left-skewed bimodal, and two symmetrical bimodal distributions with varying modal proximities. Marshall and Serlin calculates both the Huynh estimates of rho and kappa, and the

Subkoviak estimate of rho for all 15 (5 distributions x 3 lengths) tests.

Huynh's rho estimates reflected the modes in unimodal, but not bimodal, distributions. That is, the rho estimates were at their minimum value when the mode and standard converged. With the assumption of a beta distribution for Huynh's estimation procedure, this is unsurprising. In contrast, Subkoviak's rho estimates, not based on the beta-binomial assumption, performed similarly with unimodal and bimodal distributions, reflecting the modes in both cases.

That Huynh's kappa estimates measured something very different from the rho estimates is unsurprising in light of the different formulae and role of chance in the two coefficients. Kappa responded to the shape of distributions just as rho had, but in a manner opposite of rho. Kappa was at its maximum value, rather than minimum as with rho, when the standard and mode converged in both skewed and normal distributions.

Overall, research points to the impact of both test length and distribution shape on the behavior of rho and kappa estimates. Huynh and Subkoviak estimations of rho yield different patterns of over- and underestimation, though both deviated very little from the two-form estimates with a normally distributed test of 30 or 50 items.

Several important questions are as yet unanswered regarding one-form estimates of rho and kappa. Research has not been conducted with tests of length 50+, nor has the impact of distribution shape on tests of over 20 items been investigated. As many of the criterion-referenced tests used by school districts consist of more than 20 items, and distribution shapes, though generally unimodal, can vary greatly, it is important to assess the behavior of the estimates under these conditions. In addition, no research has been reported using Subkoviak's kappa estimation.

PURPOSE

The current investigation included two studies comparing the behavior of Huynh and Subkoviak estimates of rho and kappa coefficients. Study I involved simulated data, and Study II used actual test data.

In Study I, data were generated to simulate nine distributions of test scores, with 3 test lengths (25, 50, and 75 items) and 3 shapes (normal, and two degrees of skewness). Three standards for designating mastery (70%, 80% and 90% of items correct) were applied to each of the nine distributions, yielding 27 tests. Estimates of both rho and kappa estimates, as proposed by Huynh and Subkoviak, were calculated for each of the 27 distributions.

In Study II, estimates of rho and kappa proposed were calculated using data from three tests given by a large school district to assess mastery of curricular objectives. The tests, all composed of over 75 items, yielded three distinct distribution shapes varying in degree of skewness. Three standards were set at 70%, 80%, and 90% of items correct for each test. Subtests of 25, 50 and 75 items were extracted from each test by random selection and the three standards applied. Rho and kappa estimates were calculated for the nine original tests, and the 27 subtests which paralleled those of Study I.

STUDY I: SIMULATION STUDY

Data Generation

Nine distributions were simulated using the Aherns and Dieter (1974) algorithm for beta parameters. Each distribution consisted of 2500 nonzero values representing the number of students who hypothetically took each test. Test lengths of 25, 50 and 75 items, and three distribution shapes (normal, left skewed with low kurtosis, and left skewed with high kurtosis) were specified to yield nine distinct distributions. Appendix A offers a discussion of beta distributions, alpha and beta parameters, and recent research on the beta-binomial model. Table 6 displays the alpha and beta parameters used to generate the nine distributions and resultant test statistics.

Distributions

Data were generated in three distributional shapes: normal, left-skewed with low kurtosis and left-skewed with high kurtosis. Table 6 displays the statistics for these distributions.

These three shapes are representative of those seen in criterion-referenced tests used by school districts. Educational Testing Service notes a trend in Basic Skills Assessment tests toward normal and/or left-skewed

Table 6. Statistics of simulated tests

Distribution	Number of items	Mean	Standard deviation	Skewness	Kurtosis	Alpha	Beta	Mode
Normal	25	12.00	3.80	.02	-.15	4.0	4.0	12(48%)
	50	24.90	7.87	-.04	-.33	4.0	4.0	25(50%)
	75	36.96	11.61	.04	-.26	4.0	4.0	37(49%)
Left-skewed (low)	25	16.36	3.42	-.48	.09	6.0	3.0	17(68%)
	50	33.18	6.86	-.39	-.14	6.0	3.0	35(70%)
	75	50.19	10.56	-.49	.03	6.0	3.0	53(71%)
Left-skewed (high)	25	19.72	2.84	-.92	.82	8.0	2.0	21(84%)
	50	40.11	5.88	-.99	.82	8.0	2.0	43(86%)
	75	60.51	8.35	-.97	1.04	8.0	2.0	65(87%)

distributions. The Des Moines Community Schools have found normal and both right- and left-skewed distributions in district-wide tests of curricular objectives which are/can be used as criterion-referenced tests. Although distribution shapes vary, it is rare to see a bimodal distribution for criterion-referenced tests.

These distributions also allow for comparisons with findings of Subkoviak (1978) and Marshall and Serlin (1979). Subkoviak calculated rho coefficients for an (apparently) normal distribution, and Marshall and Serlin calculated rho and kappa for both normal and left-skewed distributions.

Examinees

Data were simulated representing 2500 nonzero scores for each test. This is a typical number of examinees on tests given on a comprehensive basis in large school districts and is comparable to the number of examinees in Study II.

Test Lengths

Tests were specified by lengths of 25, 50, or 75 items. Many criterion-referenced tests used on a semester or annual basis are within this range of items. Although many mastery tests used in the classroom are shorter than 25 items, the focus herein is on criterion-referenced tests given on a

districtwide basis and with which decisions are based on composite rather than subtests scores. Comparison with Subkovick (1978) calculations of rho estimates with 30 and 50 item tests is made possible.

Standards

Three standards were applied individually to each of the nine distributions, yielding a total of twenty-seven simulated tests. Standards used were 70%, 80%, and 90% of items correct, as shown in Table 7. The rationale for these standards lies in the current focus on districtwide tasting; with district development of tests and setting of standards,

Table 7. Mastery standards for simulated data

Test length	Standards		
	70% Items correct	80% Items correct	90% Items correct
25	17[a]	20	22[a]
50	35	40	45
75	52[a]	60	67[a]

[a]Score rounded downward from (score + 0.5).

it is usual for tests to be of a difficulty level such that most examinees answer most of the items correctly. In general, only a minority (10-20%) of examinees are classified as nonmasters with districtwide tests.

Rho and Kappa Estimates

Subkoviak and Huynh estimates of coefficients rho and kappa were calculated for each of the twenty-seven simulated test distributions. Estimates were calculated from computer programs developed by the respective authors (Subkoviak, 1978; Huynh & Saunders, 1980) and adapted by the present researcher.

STUDY II: ACTUAL PLAN

A parallel investigation using actual rather than simulated data was also conducted. Three tests administered by a large school district provided test scores for the calculation of Huynh and Subkoviak estimations of rho and kappa. The three tests yielded three distinct distributions, one appearing approximately normal, and two left-skewed. Subtests of 75, 50, and 25 items were drawn by a random sampling process from each test, and standards of 70%, 80%, and 90% of items correct were applied. This procedure created 27 tests which paralleled those in the simulation study. Estimates of rho and kappa were calculated for these 27 test distributions as well as for nine tests created by applying the three standards to the full-length tests.

The value of examining the behavior of the rho and kappa estimates with actual data is clear. Distributions of scores are not necessarily of the beta family of distributions. In the case of Huynh estimates, it is assumed that score distributions are within the beta family (see Appendix A); coefficient behavior may differ in distributions which deviate from this assumption. Calculation of estimates with data from tests actually employed by schools provides this needed real life information.

Data Collection

Scores on three tests developed and administered by the Des Moines (Iowa) Independent School District provided the data for this investigation. The objectives-based tests were developed, pilot-tested and administered by the Des Moines Independent School District as part of an ongoing curriculum evaluation program. None of the tests was used as a criterion-referenced test at the time of administration, although curriculum specialists and individual teachers were encouraged to evaluate individual students on the basis of the test, as well as using it for evaluation of their own teaching. One of the tests, biology, was used as all or part of the students' final examination in that course.

Two tests, mathematics and geography, were administered to all seventh-grade students, while the third test, biology, was administered to all students, predominantly tenth-graders, enrolled in that course. All tests were intended and used to evaluate mastery of the core objectives of the respective courses. Table 8 displays the test statistics for the three tests.

Table 8. Statistics of actual tests

Distribution	Number of items	Mean	Standard deviation	Skewness	Kurtosis	Alpha	Beta	Mode
Math	25	14.58	4.54	-.02	-.57	5.3	3.8	13(52%)
(N=2282)	50	27.03	9.03	.25	-.71	4.2	3.6	25(50%)
	75	39.92	13.32	.28	-.71	4.1	3.6	28(37%)
	95	48.24	16.57	.38	-.62	4.0	3.9	37(39%)
Geography	25	16.28	4.40	-.31	-.52	5.8	3.1	16(64%)
(N=2160)	50	32.83	8.76	-.30	-.64	4.9	2.6	31(62%)
	75	48.27	12.77	-.25	-.67	5.0	2.8	46(61%)
	80	52.43	13.44	-.31	-.60	5.1	2.7	51(64%)
Biology	25	17.18	4.07	-.41	-.25	7.3	3.3	17/20(61%/80%)
(N=1216)	50	33.39	7.64	-.30	-.40	7.0	3.5	
	75	51.68	10.91	-.44	-.26	7.3	3.3	
	88	60.89	12.54	-.46	-.28	7.5	3.3	

Subtests

Subtests of 25, 50, and 75 items were drawn from the full-length tests. Items on the full-length tests were randomly deleted to create tests of 75 items. From the 75 remaining items, 25 were randomly deleted to create 50 item tests, and 25 items from these forms were randomly deleted to create 25 item tests. Thus, all versions of the shorter tests were composed only of items contained in the longer tests. While this procedure led to nonindependence of tests, it maximized the similarity of distribution shapes across test lengths. The test statistics for the nine subtests (three tests x 3 lengths) are displayed in Table 8, above, along with those for the full-length tests.

Examinees

Test scores for all students who were enrolled in the respective courses and who took the tests were included in the analysis. The number of examinees were 2282, 2160, and 1216 for the math, geography, and biology tests, respectively. The smaller number of examinees for the high school biology test is accounted for by the elective nature of that course; the math and geography are required courses for all seventh grade students.

Distributions

The distribution of scores on the math test appeared approximately normal, while the geography and biology tests were left-skewed. Alpha and beta parameters were calculated for full-length tests as well as subtests. Parameters for the different lengths of the same tests vary slightly due to the random selection of items in subtests. All parameters are displayed in Table 8.

Although there is no statistical procedure which provides a cogent test for the goodness-of-fit of the beta-binomial model to data, a descriptive technique was employed to provide a general indication of whether test distributions were in the beta family. (Appendix A offers a discussion of the beta-binomial model and the difficulty of evaluating its goodness-of-fit to data.) Alpha and beta parameters for the 12 test distributions shown in Table 8 were used to generate between distributions. Frequencies yielded by the beta distributions were compared with observed frequencies of the comparable tests. The maximum discrepancy between observed and expected frequencies (D_{max}), and chi-square goodness-of-fit statistics were calculated (see Table 9).

In every case, the chi-square statistic was significant at the .01 level, signifying that observed departed significantly from expected frequencies. However, this

Table 9. Discrepancies between expected and observed frequencies

Test (items)	D_{max}(%)	df	χ^2
Math			
92	29.5 (.012)	72	779.9**
75	37.8 (.016)	59	495.5**
50	58.1 (.030)	42	507.3**
25	90.2 (.039)	21	3325.2**
Biology			
88	19.4 (.015)	53	121.9**
75	19.2 (.015)	46	160.3**
50	43.1 (.035)	33	241.6**
25	75.6 (.062)	17	745.3**
Geography			
80	35.2 (.020)	58	219.7**
75	29.5 (.013)	56	242.0**
50	61.5 (.028)	39	359.9**
25	120.6 (.055)	20	1103.1**

**$p<.01$.

appears to be an artifact of the necessary procedure used to group categories at the tails of the distributions (see Appendix B).

Standards

Three mastery standards were applied to the three full-length tests and to the nine subtests. Standards paralleled those used in Study I: 70%, 80%, and 90% of items correct. As stated above, these are commonly found mastery standards for criterion-referenced tests given on a

districtwide basis. Table 10 displays the mastery standards for all full-length tests and subtests. Note that all standards are rounded down to the nearest whole number.

Table 10. Mastery standards for actual test data

Test	Number of items	Standards		
		70%	80%	90%
Subtests	25	17[a]	20	22[a]
Subtests	50	35	40	45
Subtests	75	52[a]	60	67[a]
Geography	80	56	64	72
Biology	88	61[a]	70[a]	79[a]
Math	95	66[a]	76	85[a]

[a]Rounded down to nearest whole number.

Rho and Kappa Estimates

Huynh and Subkoviak estimates of rho and kappa were calculated for all tests and subtests with all three levels of mastery.

RESULTS

Results of Study I are discussed below, followed by results of Study II. Both sections begin with presentations of all estimates of rho and kappa calculated and are organized in parallel fashion.

Within each section, rho and kappa estimates are discussed separately, beginning with a general description of findings for respective estimates, and proceeding to the impact of test lengths, distribution shapes, and standards on coefficients. Brief summaries conclude each discussion of rho and kappa.

As no acceptable statistical means of comparison are available, estimates are evaluated descriptively.

Other than Tables 11, 18 and 19, which present an overview of all coefficients, all rho and kappa estimates are rounded to the hundredths place for ease in reading.

Study I: Simulated Data

Rho estimates

Table 11 displays rho estimates calculated by the Huynh and Subkoviak procedures for all simulated test distributions. Huynh estimates of rho ranged from .678 to .996, with a median value of .87; Subkoviak estimates ranged from .740 to .994, with a median value of .88. As can be seen, Huynh and

Table 11. Rho and kappa estimates for simulated data (N=2500)

	Number of items	Standard	Rho estimates		Kappa estimates	
			Huynh	Subkoviak	Huynh	Subkoviak
Normal	25	70%	.854	.856	.332	.481
		80%	.964	.942	.203	.388
		90%	.992	.979	.107	.291
	50	70%	.903	.895	.538	.633
		80%	.966	.961	.427	.537
		90%	.995	.991	.251	.287
	75	70%	.922	.926	.612	.668
		80%	.976	.972	.498	.590
		90%	.996	.994	.329	.455
Left-skewed (low)	25	70%	.684	.741	.368	.482
		80%	.793	.794	.306	.433
		90%	.918	.890	.206	.371
	50	70%	.788	.802	.572	.603
		80%	.852	.859	.518	.584
		90%	.957	.945	.459	.459
	75	70%	.834	.833	.668	.667
		80%	.878	.880	.623	.642
		90%	.956	.954	.497	.581
Left-skewed (high)	25	70%	.834	.852	.296	.478
		80%	.678	.748	.344	.479
		90%	.715	.740	.308	.430
	50	70%	.874	.890	.550	.632
		80%	.796	.826	.579	.634
		90%	.822	.812	.526	.541
	75	70%	.903	.912	.612	.663
		80%	.839	.842	.644	.668
		90%	.845	.842	.601	.609

Subkoviak estimates deviate little from one another (the largest difference being .06). Both Huynh and Subkoviak estimates are reasonably high relative to values expected for reliability coefficients.

Table 11 also indicates that Huynh and Subkoviak estimates differ with test length, shape of distribution, and standard. The effects of these variables on rho estimates are discussed below.

Test lengths Given the same distribution and standard, longer tests yield higher rho coefficients without exception. Median values for Huynh and Subkoviak estimates by test length are shown in Table 12. As shown, Subkoviak estimates are slightly higher than Huynh estimates, but this difference is not great enough to be of practical importance.

Table 12. Median rho estimates for simulated data by test length

Test length	Huynh estimates	Subkoviak estimates
25 items	.83	.85
50 items	.87	.89
75 items	.90	.91

Distribution shapes In every case, at each standard and test length, both Huynh and Subkoviak estimates for the normal distribution are higher than those for comparable standards and test lengths of the skewed distributions. This is reflected by the medians for each distribution reported in Table 13.

Table 13. Median rho estimates for simulated data by distribution shape

	Huynh estimates	Subkoviak estimates
Normal	.96	.96
Left-skewed (low)	.87	.86
Left-skewed (high)	.83	.84

Standards No consistent pattern was seen regarding the impact the standard across all test lengths and distribution shapes.

Standards and distribution shapes Table 14 displays the median rho estimates by standard and distribution shape. Huynh estimates are followed by Subkoviak estimates in each case.

The normal and left-skewed (low) distributions showed similar patterns: as the standard increased, the rho coefficients increased. The left-skewed (high) distribution

Table 14. Median rho estimates for simulated data by standard and distribution

Standards	Normal (49%)[a]	Left-skewed (low) (70%)	Left-skewed (high) (86%)
70%	.90/.89[b]	.78/.80	.87/.89
80%	.96/.96	.85/.85	.79/.82
90%	.99/.99	.95/.94	.82/.81

[a]Distribution mode.

[b]Huynh estimate/Subkoviak estimate.

did not reveal the same pattern: here, both Huynh and Subkoviak coefficients were at their maximum at the lowest standard. Figures 2-4 depict these relationships between standards and distributions for each test length graphically.

In all three distributions, the behavior of both Huynh and Subkoviak estimates of rho reflect the proximity of the mode to the standard. That is, the rho coefficients were at their minimum observed value when the standard was near the mode of the distribution. The modes for normal, left-skewed (low), and left-skewed (high) distributions were within .01 of 49%, 70%, and 86%, respectively. (Variance in modes was due to rounding at different test lengths.)

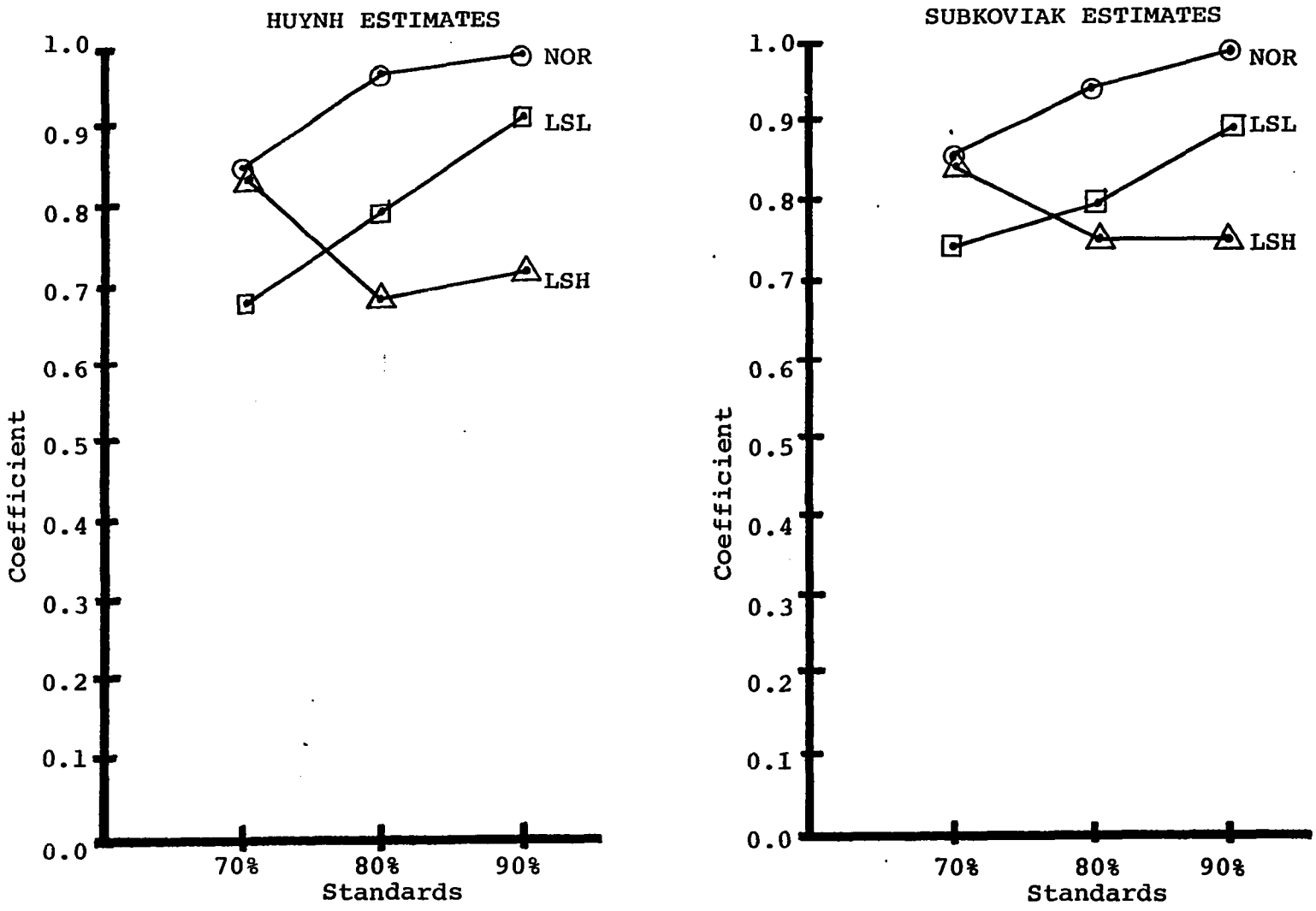

Figure 2. Rho estimates for simulated 25 item tests

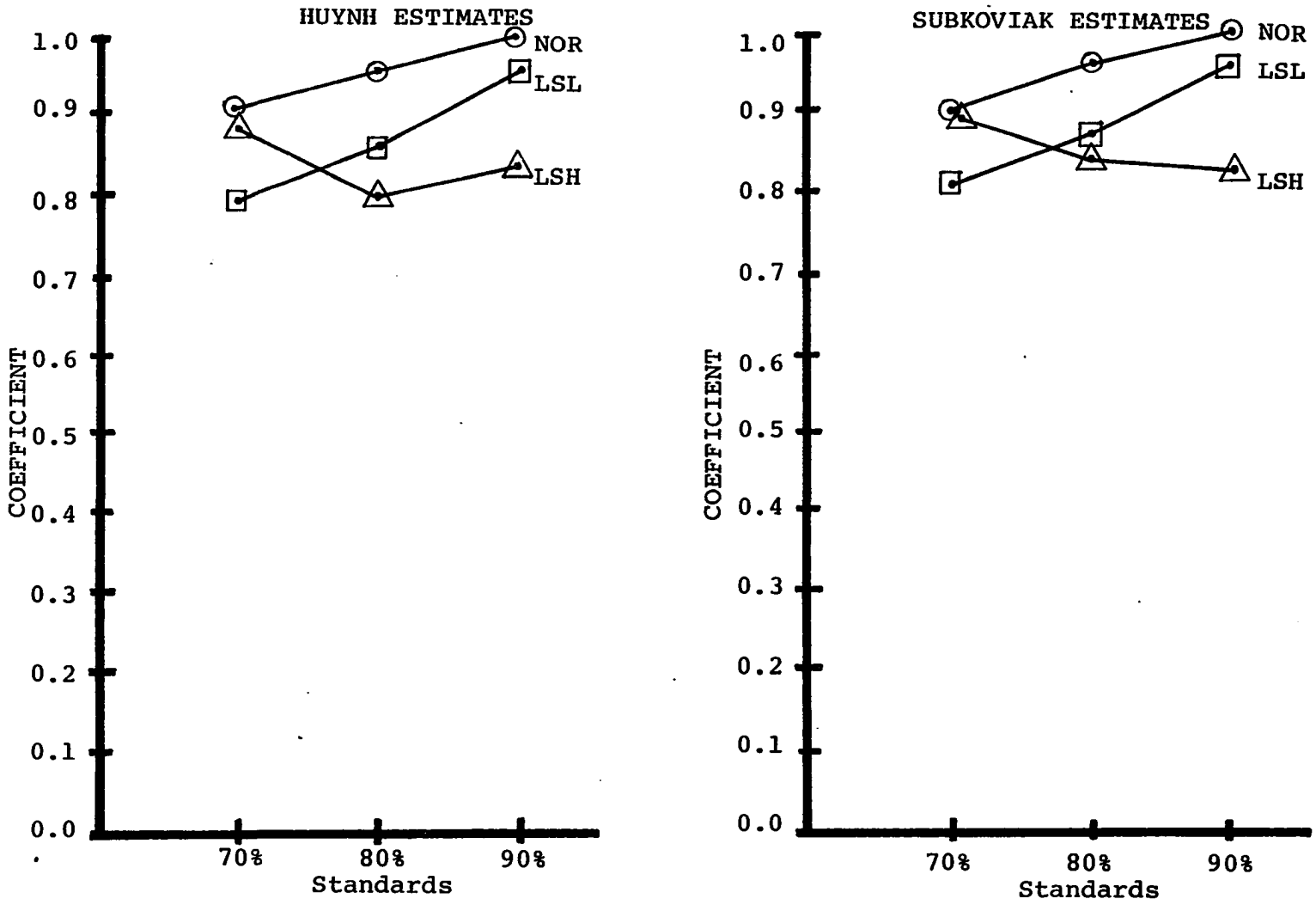

Figure 3. Rho estimates for simulated 50 item tests

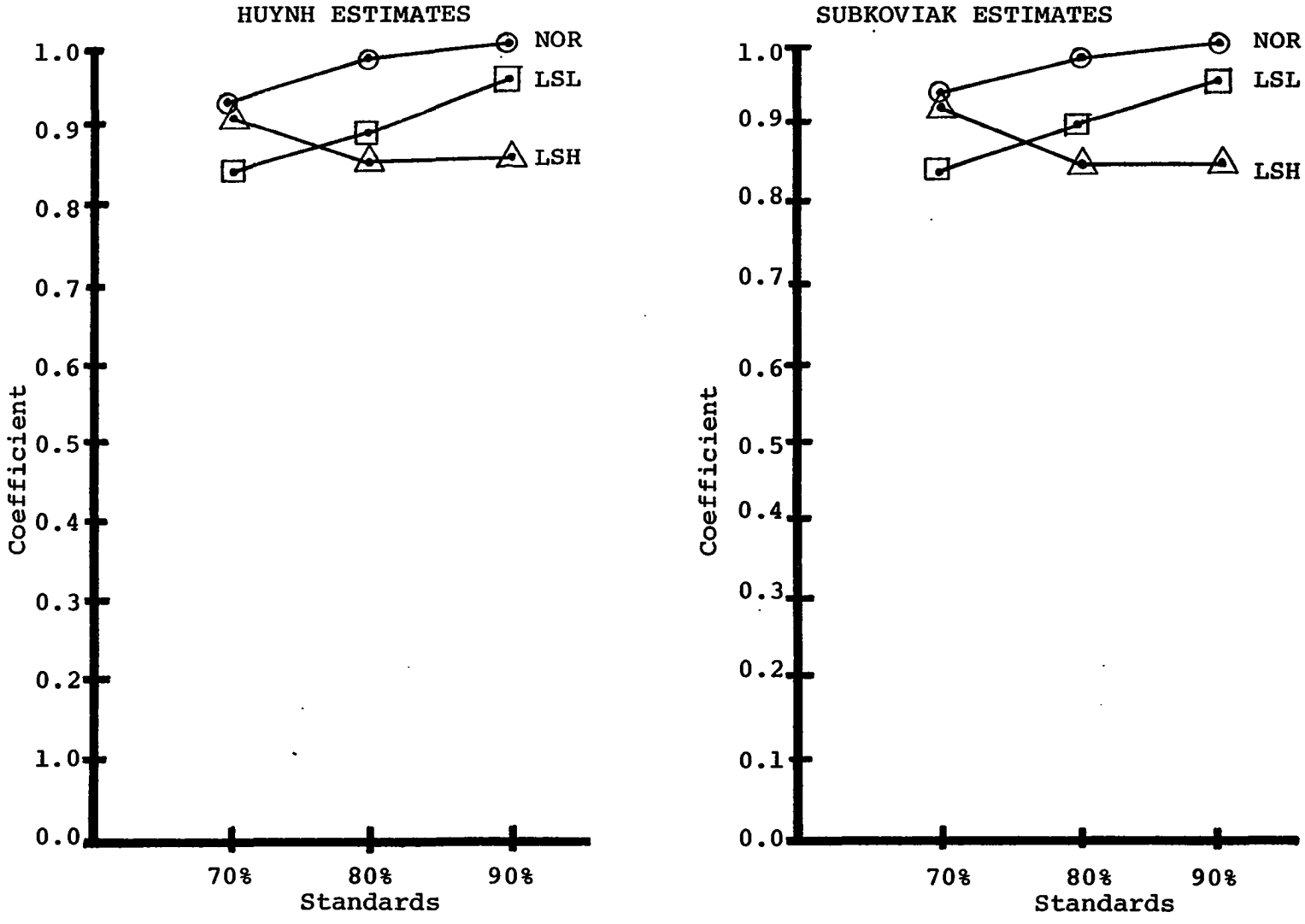

Figure 4. Rho estimates for simulated 75 item tests

Summary Overall, both Huynh and Subkoviak estimates of rho with simulated test data were similar and reasonably high. Longer tests yielded higher estimates in all cases; the normal distribution displayed higher estimates than skewed distributions. Test lengths and the interactions between standards and distribution shapes had an impact on coefficients: estimates were at their lowest observed value when the standards were at or near the distribution modes in all cases.

When the standards and modes coincide, a small difference in form A and B scores (e.g., one point) for examinees leads to inconsistency in categorization for the largest number of examinees. When the standards are far from the modes, a small difference in scores for examinees leads to less inconsistency in categorization because fewer examinees are near this critical area of the standard (shows most inconsistency) when the standards and needs converge, and displays increasing values as the standards and modes diverge. Thus, rho is at its lowest value of the standard.

Kappa estimates

Kappa estimates for simulated data are displayed in Table 11. As expected, kappa estimates were lower than rho estimates. Huynh estimates ranged from .107 to .668, with a

median value of .49; Subkoviak estimates ranged from .291 to .668, with a median value of .54. Subkoviak estimates were higher than Huynh estimates by at least .10 for comparable distributions, test lengths and standards in 15 of the 27 cases, and by at least .02 in 25 of the 27 cases.

Test lengths Given the same standard and distribution, longer tests resulted in higher coefficients without exception. For comparable lengths, Subkoviak estimates were higher than Huynh estimates. Median values by test length are displayed in Table 15.

Table 15. Median kappa estimates for simulated data by test length

Test length	Huynh estimates	Subkoviak estimates
25 items	.30	.43
50 items	.52	.60
75 items	.61	.64

Distribution shapes Median values for coefficients by distribution are shown in Table 16. As can be seen, the normal distribution yielded the lowest kappa estimates. Subkoviak estimates are consistently higher than Huynh estimates (by from .05-.15) in all distributions.

Table 16. Median rho estimates for simulated data by distribution shape

	Huynh estimates	Subkoviak estimates
Normal	.33	.48
Left-skewed (low)	.49	.58
Left-skewed (high)	.58	.63

Standards No consistent pattern was discernible regarding the impact of standards on kappa estimates. An interaction, however, of standards with distribution shapes was evident and is discussed below.

Standards and distribution shapes Table 17 displays the median values for Huynh and Subkoviak estimates of kappa by distribution and standard. Huynh estimates are followed by Subkoviak estimates in each case. Graphic representations follow in Figures 5-7.

The normal and left-skewed (low) distributions showed a similar behavior in regard to the standard: as the standard increased, the kappa estimates decreased. This, it will be noted, was the opposite of rho estimates where coefficients increased as the standard increased. The left-

Table 17. Median kappa estimates by standard and distribution

Standards	Normal (49%)[a]	Left-skewed (low) (70%)	Left-skewed (high) (86%)
70%	.53/.63[b]	.57/.60	.55/.63
80%	.42/.53	.51/.58	.57/.63
90%	.25/.28	.36/.45	.52/.54

[a]Distribution mode.

[b]Huynh estimates/Subkoviak estimates.

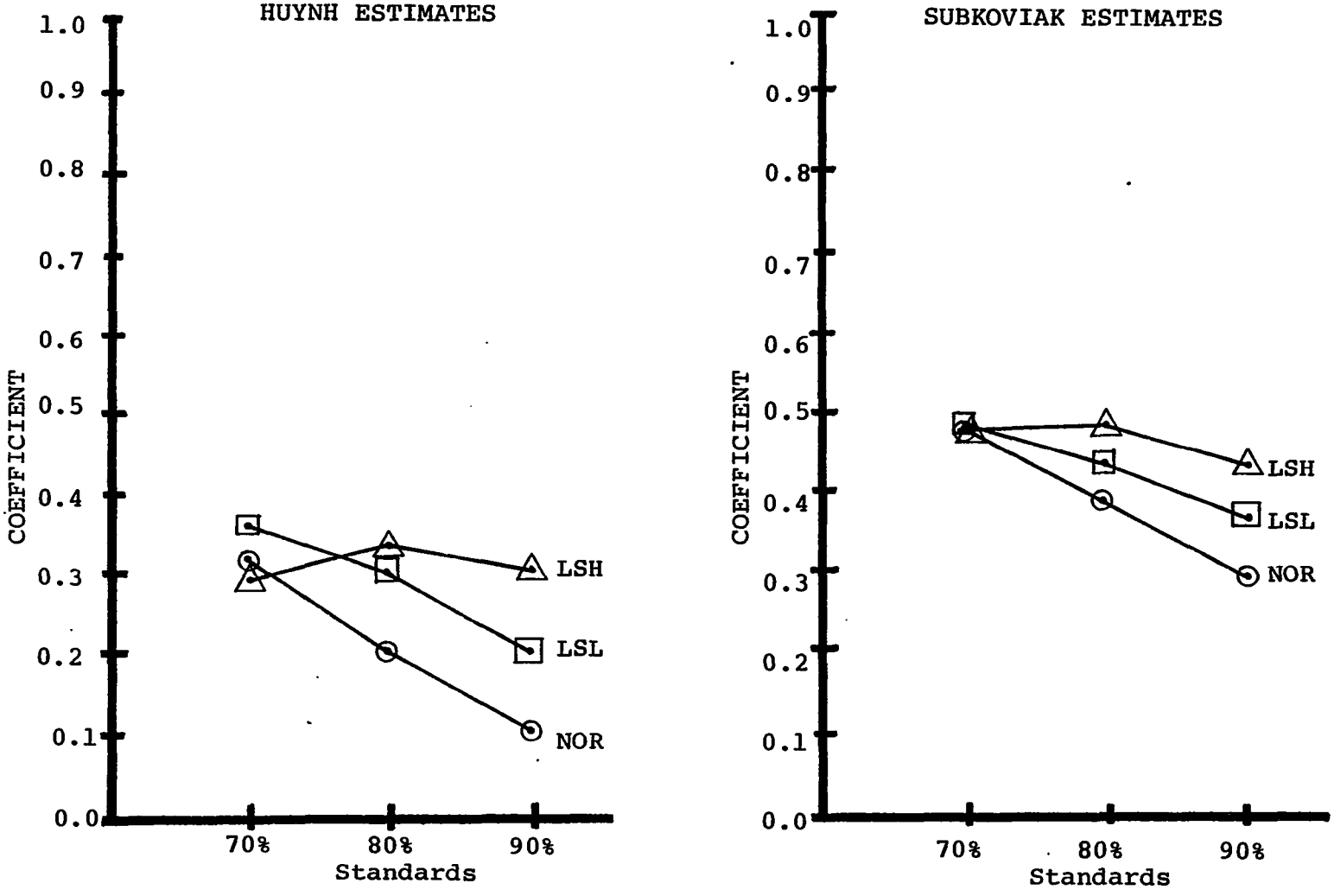

Figure 5. Kappa estimates for simulated 25 item tests

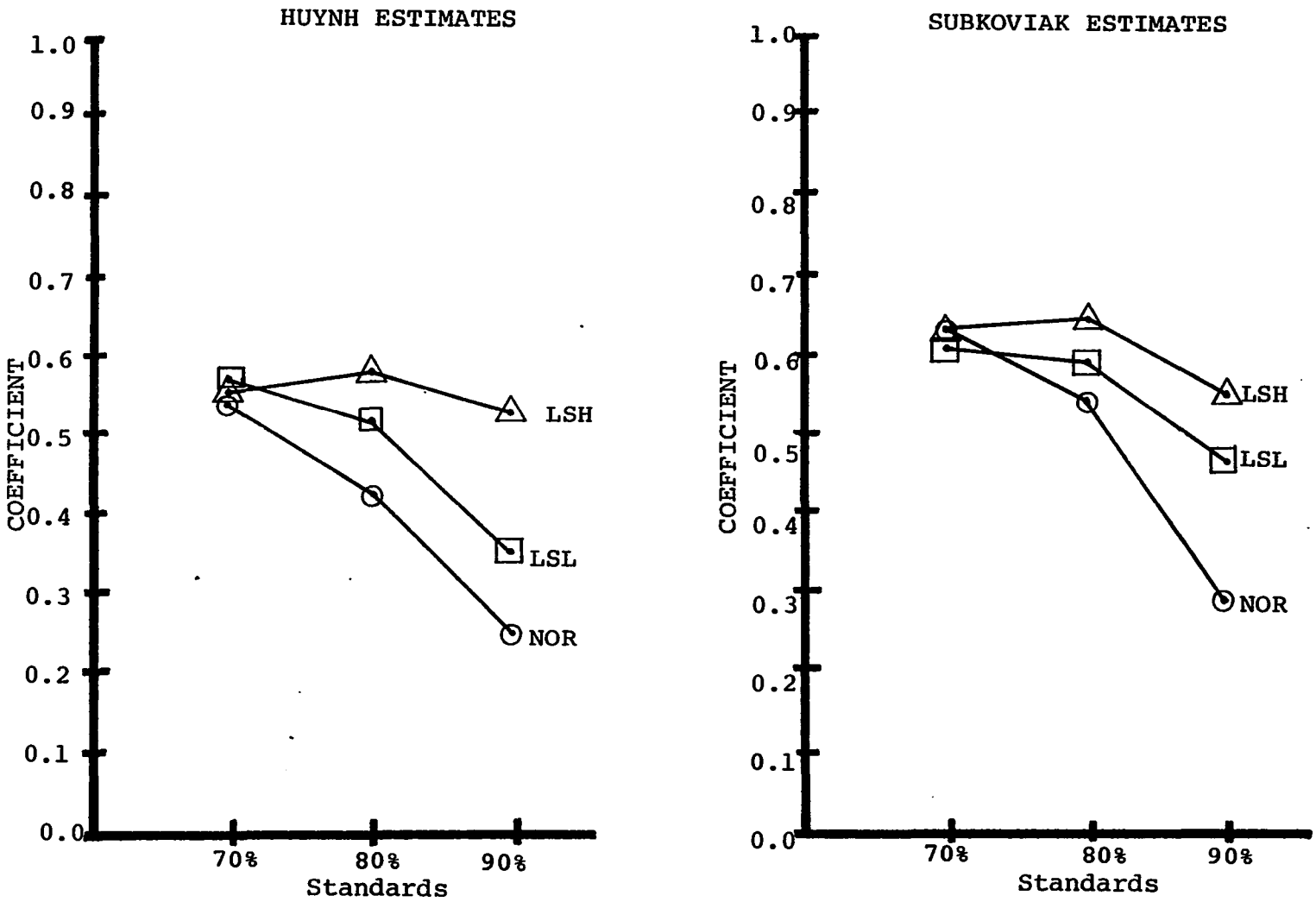

Figure 6. Kappa estimates for simulated 50 item tests

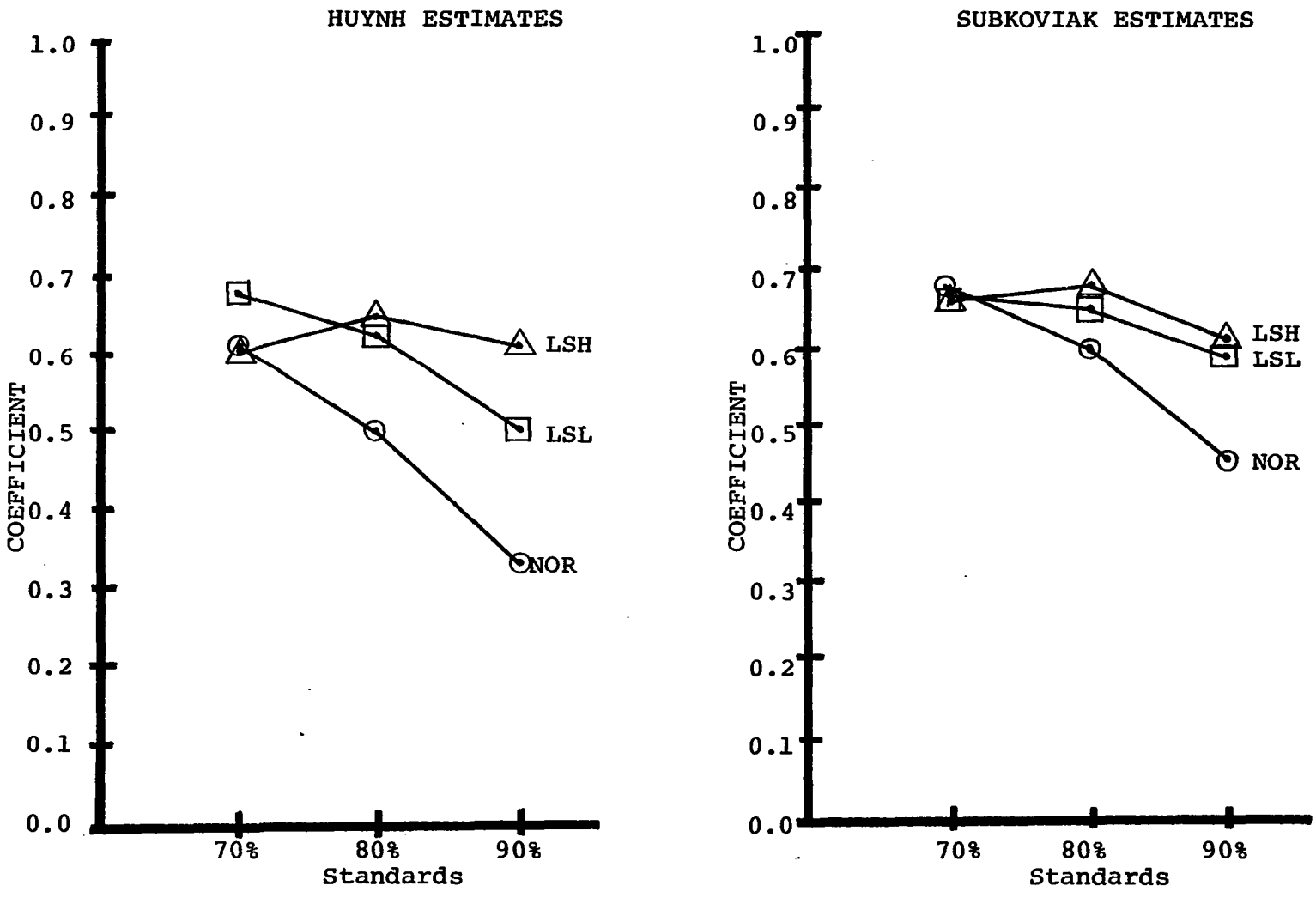

Figure 7. Kappa estimates for simulated 75 item tests

skewed (high) distribution did not follow this pattern; it yielded a maximum coefficient at the 80% standard and minimum at the 90% standard.

As mentioned above, the modes for the normal, left-skewed (low) and left-skewed (high) were 49%, 70%, and 86%, respectively. Kappa coefficients, thus, reflected the distribution modes in that they were at their maximum observed value when the standard approached the mode. This was the opposite of the rho coefficients, which were at their minimum value when the standard was near the mode.

Summary Kappa coefficients were lower than rho coefficients at comparable test lengths, distribution shapes and standards. This was not unexpected, as both Huynh and Subkoviak estimates of kappa take the effect of chance into account and rho estimates do not.

Increasing test length increased the magnitude of estimates, but even with the longest tests, estimates were not at a level considered acceptably high for reliability coefficients. Both Huynh and Subkoviak estimates reflected the mode in that estimates were at their maximum value when standards converged with distribution modes.

Study II: Actual Data

Study II used scores from three tests administered by a school district, as well as subtests of 25, 50 and 75 items created from each of the tests. Math scores appeared quite normally distributed, and the geography and biology scores were increasingly left-skewed.

None of the test distributions in Study II was strictly of the beta-binomial family according to goodness-of-fit test performed (see Appendix B). However, serious questions remain regarding the persuasiveness of these findings (see Appendix A).

Rho estimates

Table 18 displays rho estimates calculated by Huynh and Subkoviak procedures for all subtests of math, geography and biology tests. Huynh estimates for the 25-75 item tests range from .763 to .986, with a median value of .86; Subkoviak estimates range from .795 to .977, with a median value of .87. As can be seen, Huynh and Subkoviak estimates deviate only slightly from one another, the largest difference being .03. In 16 cases at 70% and 80% standards, Subkoviak estimates are slightly higher than Huynh estimates, but not large enough to be of practical importance.

Table 19 displays estimates for the full length tests. Note that in reporting medians in Table 19 and all following

Table 18. Rho and kappa estimates for actual data

	Number of items	Standard	Rho estimates		Kappa estimates	
			Huynh	Subkoviak	Huynh	Subkoviak
Math (N=2282)	25	70%	.782	.814	.527	.600
		80%	.862	.877	.458	.590
		90%	.931	.923	.364	.544
	50	70%	.876	.902	.642	.733
		80%	.932	.930	.571	.674
		90%	.980	.969	.426	.575
	75	70%	.900	.920	.701	.774
		80%	.950	.946	.630	.708
		90%	.986	.977	.498	.613
Geography (N=2160)	25	70%	.770	.805	.540	.610
		80%	.808	.829	.500	.583
		90%	.880	.874	.425	.511
	50	70%	.841	.857	.680	.713
		80%	.867	.883	.648	.705
		90%	.932	.923	.547	.591
	75	70%	.886	.878	.728	.752
		80%	.894	.907	.691	.740
		90%	.951	.943	.596	.622
Biology (N=1216)	25	70%	.763	.795	.509	.582
		80%	.776	.807	.483	.568
		90%	.851	.854	.412	.510
	50	70%	.813	.830	.626	.660
		80%	.852	.860	.586	.627
		90%	.941	.937	.460	.573
	75	70%	.845	.854	.688	.706
		80%	.867	.868	.657	.670
		90%	.939	.936	.554	.590

Table 19. Rho and kappa estimates for actual data

	Number of items	Standard	Rho estimates		Kappa estimates	
			Huynh	Subkoviak	Huynh	Subkoviak
Math (N=2282)	95	70%	.922	.948	.716	.791
		80%	.966	.980	.641	.721
		90%	.992	.993	.502	.679
Geography (N=2160)	80	70%	.869	.878	.735	.754
		80%	.895	.905	.703	.744
		90%	.952	.945	.602	.623
Biology (N=1216)	88	70%	.854	.868	.706	.733
		80%	.873	.869	.680	.680
		90%	.948	.946	.567	.603

tables, only subtests (i.e., 25, 50 and 75 item tests) are discussed unless otherwise specified.

Test lengths Given the same distribution and standard, longer tests yield higher rho coefficients without exception. Median values for Huynh and Subkoviak estimates by test length are shown in Table 20. It is clear that at given test lengths, Subkoviak estimates are slightly higher than Huynh estimates, but not enough to be of practical importance.

Table 20. Median rho estimates for actual data by test length

Test length	Huynh estimates	Subkoviak estimates
25 items	.80	.82
50 items	.87	.90
75 items	.90	.92

Distribution shapes At all standards and test lengths, both Huynh and Subkoviak estimates for the math test are higher than for the biology and geography tests. It should be recalled that the math test scores appeared quite normally distributed, while the biology and geography distributions were clearly skewed (see Table 8). Medians for Huynh and Subkoviak estimates for each distribution are reported in Table 21. Only subtest coefficients are

Table 21. Median rho estimates for actual data by distribution[a]

	Huynh estimates	Subkoviak estimates
Math	.93	.92
Geography	.86	.88
Biology	.85	.85

[a]Includes only subtests.

included in calculations of medians.

Standards Within every test length and distribution shape, rho estimates increased with an increase in the standard. This was without exception for all subtests, as well as full length tests.

Standards and distribution shapes Table 22 displays the median rho estimates by standard and distribution. Huynh estimates are followed by Subkoviak estimates in each case. Graphic representations follow in Figures 8-10.

In all three distributions, the behavior of both Huynh and Subkoviak estimates of rho reflected the proximity of the mode to the standard. That is, the rho coefficients were at their lowest observed value as the standard neared

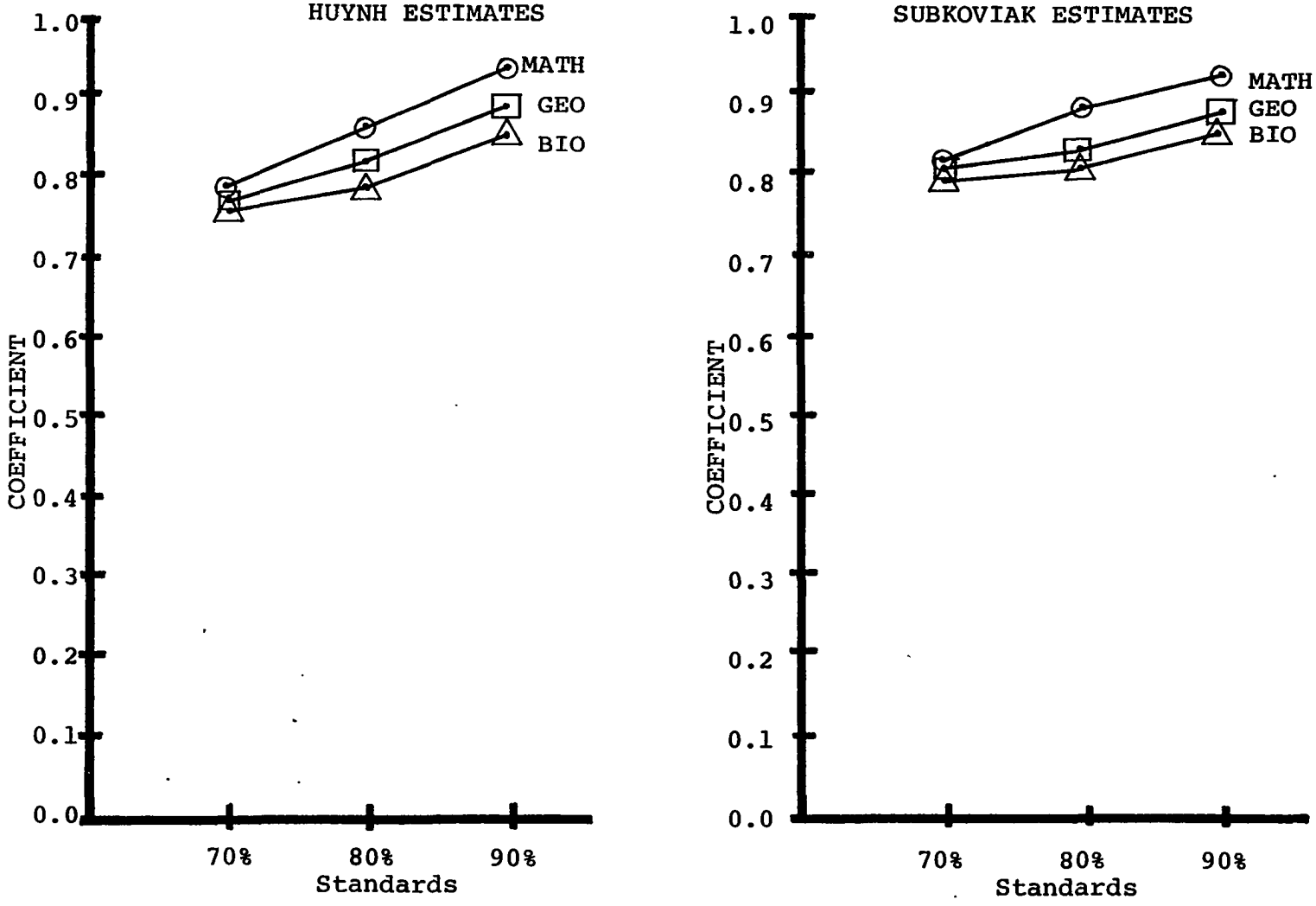

Figure 8. Rho estimates for actual 25 item tests

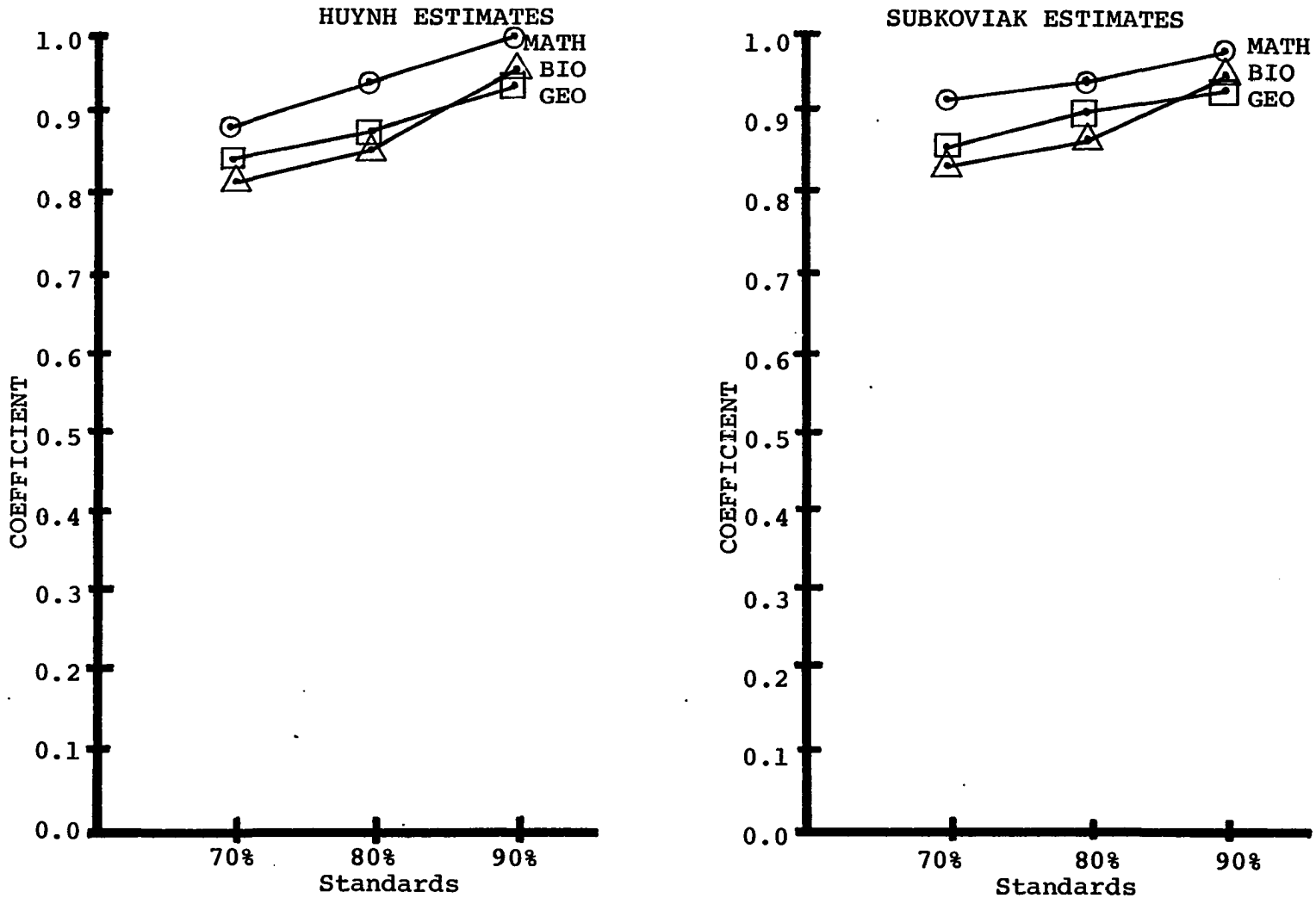

Figure 9. Rho estimates for actual 50 item tests

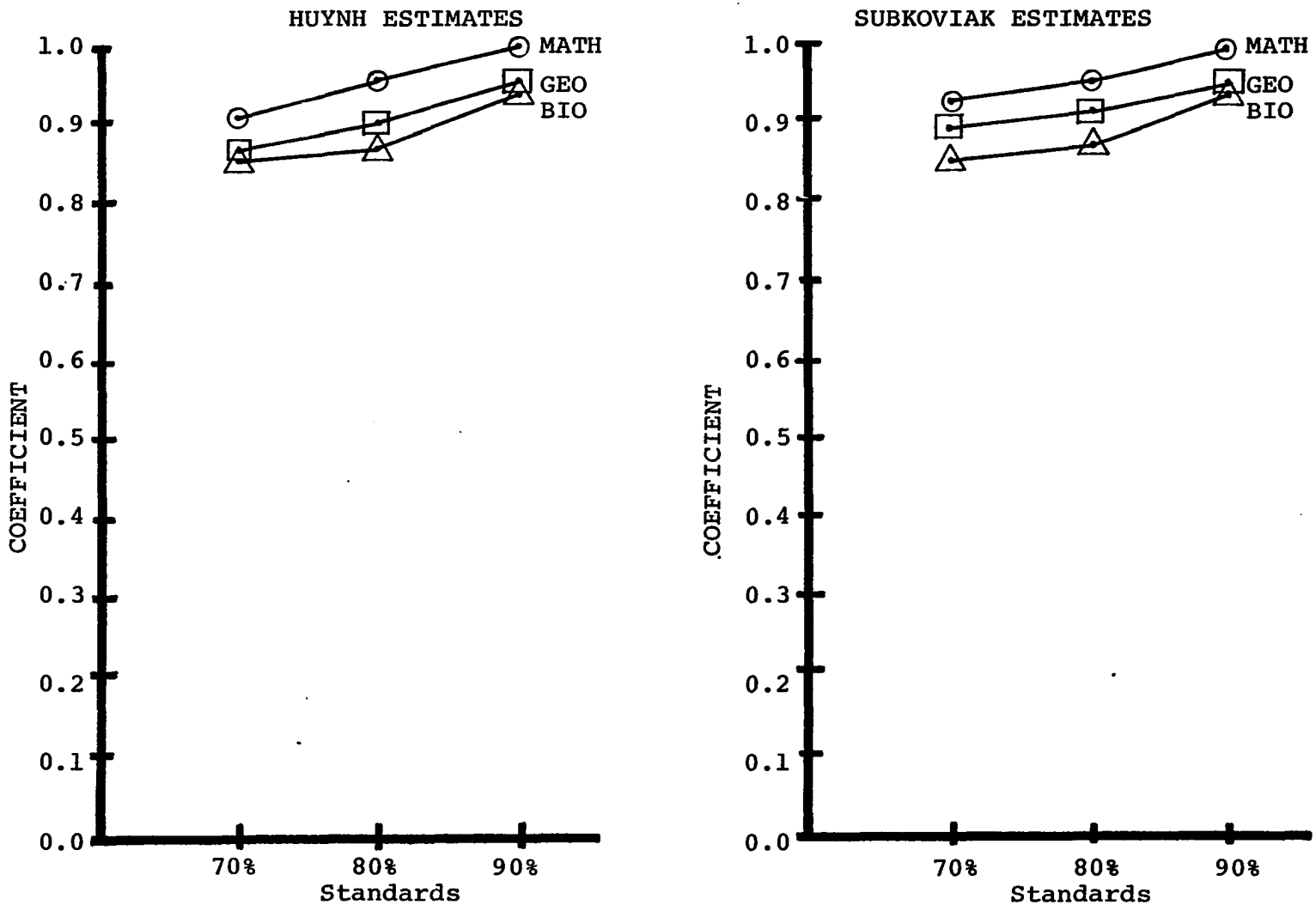

Figure 10. Rho estimates for actual 75 item tests

Table 22. Median rho estimates for actual data by standard and distribution

Standards	Math (46%)[a]	Geography (62%)	Biology (72%)
70%	.87/.90[b]	.84/.85	.81/.83
80%	.93/.93	.86/.88	.85/.86
90%	.98/.96	.93/.92	.93/.93

[a]Average distribution mode.

[b]Huynh estimate/Subkoviak estimate.

the mode of the distribution. The average modes for all math, geography, and biology subtests were 46%, 62%, and 72%, respectively. Modes for subtests varied with length due to the random selection of items (see Table 8).

Beta-binomial model Although none of the distributions was strictly of the beta family (see Appendices A and B), the behavior of Huynh estimates (which assume a beta-binomial distribution) did not differ from that expected if the distributions had been from the beta family. In fact, the behavior paralleled that of the Huynh estimates in Study I, which used data which fit the beta-binomial model. In both studies, the proximity of the distribution mode and standard yielded the lowest observed

coefficients regardless of distribution shape.

Summary Huynh and Subkoviak estimates of rho based on actual data deviated very little from one another. All coefficients were reasonably high in terms of expected values for reliability coefficients.

Coefficients increased with increased test length in all cases. Both Huynh and Subkoviak estimates behaved similarly in terms of distribution shape and location of the standard: lowest observed values were seen when the standards were near the distribution modes.

Kappa estimates

Kappa estimates for subtests of actual data are displayed in Table 18, and for full length tests in Table 19. Huynh's kappa estimates for subtests ranged from .364 to .728, with a median value of .55; Subkoviak's estimates ranged from .510 to .774, with a median value of .61. In every case, the Subkoviak estimates were higher than the Huynh estimates by at least .03.

Test lengths Given the same standard and distribution, longer tests resulted in higher kappa coefficients without exception. In addition, Subkoviak estimates were higher than Huynh estimates at every length. Median values for subtests by length are displayed in Table 23.

Table 23. Median kappa estimates for actual data by test length

Test length	Huynh estimates	Subkoviak estimates
25 items	.48	.58
50 items	.58	.66
75 items	.62	.71

Distribution shapes Median values for kappa coefficients by distribution are shown in Table 24. The geography test yielded the highest median coefficients, followed by biology and math.

Table 24. Median kappa estimates for actual data by distribution

	Huynh estimates	Subkoviak estimates
Math	.52	.61
Geography	.59	.62
Biology	.55	.59

Standards Within every test length and distribution shape, kappa coefficients decreased as standards increased. This behavior was the opposite of rho estimates, where coefficients increased with the standard.

Standards and distribution shapes Table 25 displays the median values for Huynh and Subkoviak estimates of kappa by distribution and standard. Huynh estimates are followed by Subkoviak estimates in each case. Figures 11-13 depict the relationships graphically.

Table 25. Median kappa estimates for actual data by standard and distribution

Standards	Math (46%)[a]	Geography (62%)	Biology (72%)
70%	.64/.73[b]	.68/.71	.62/.66
80%	.57/.67	.64/.70	.58/.62
90%	.42/.57	.54/.59	.46/.57

[a]Average distribution modes.

[b]Huynh estimate/Subkoviak estimate.

In all three distributions, the behavior of kappa reflected the proximity of the standard to the mode. The average modes for the math, geography and biology tests were 46%, 62%, and 72%, respectively. (Differences in subtest modes for a given test were due to random selection of items.) The highest observed value was seen at the 70% standard in all cases, including full-length tests.

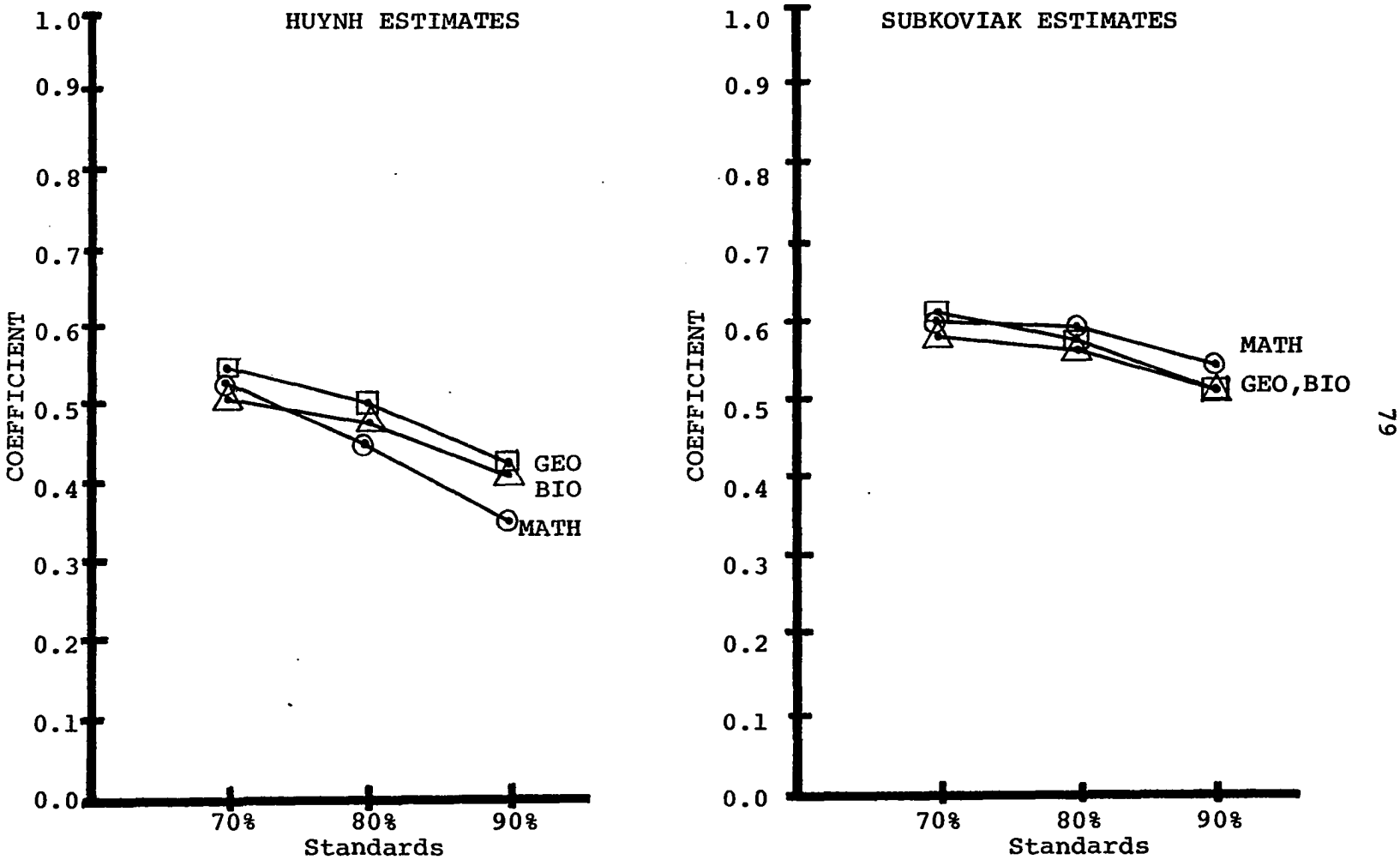

Figure 11. Kappa estimates for actual 25 item tests

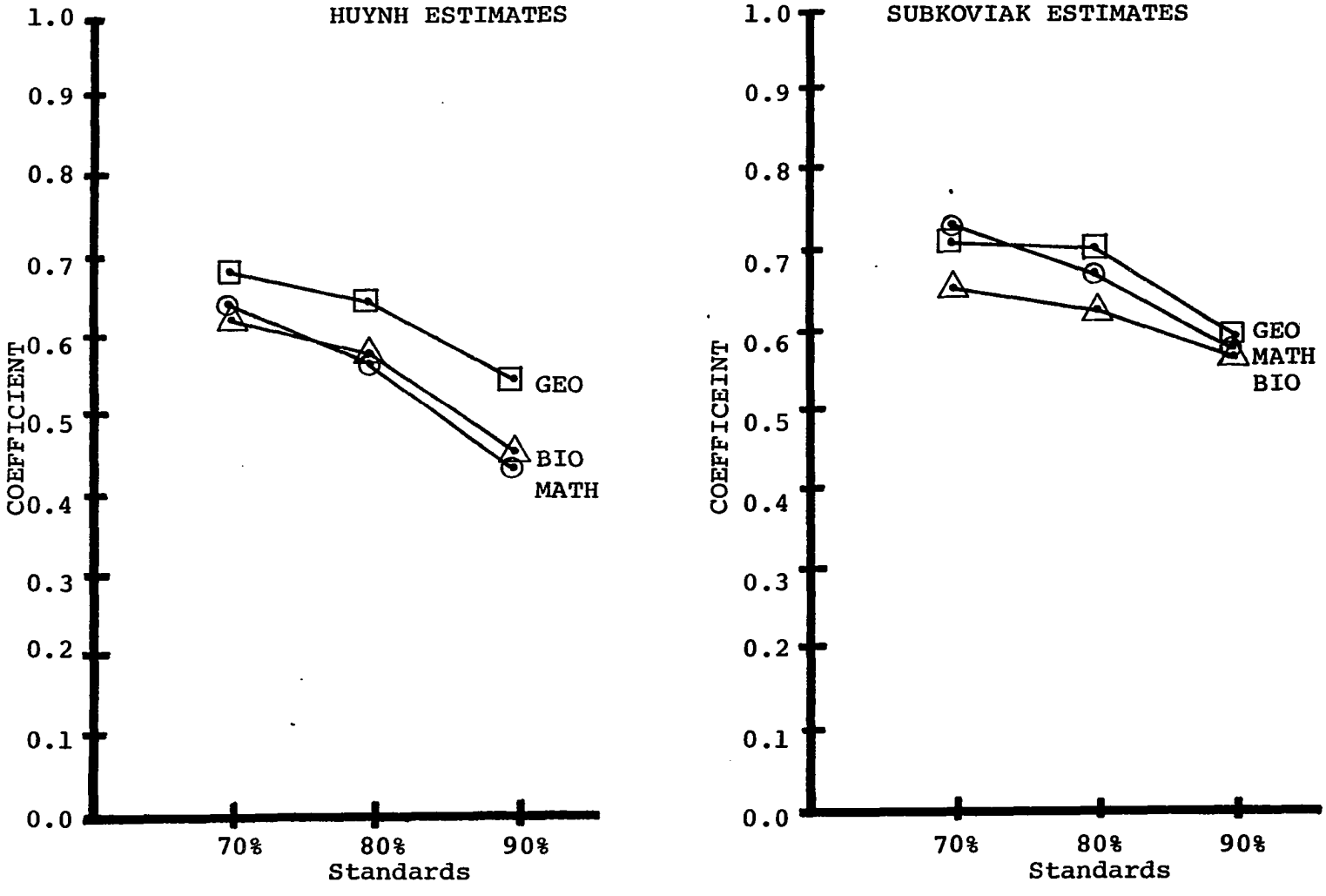

Figure 12. Kappa estimates for actual 50 item tests

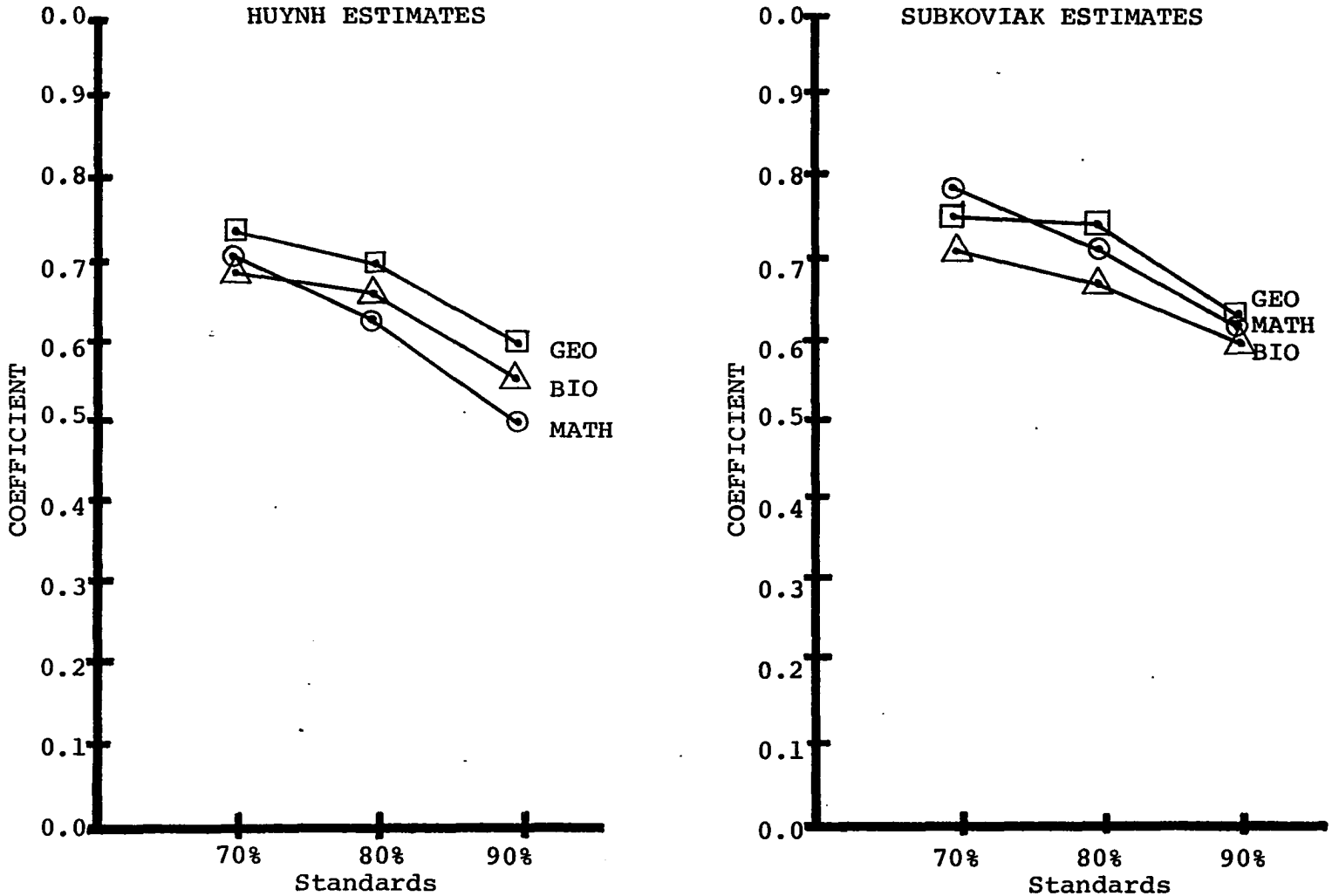

Figure 13. Kappa estimates for actual 75 item tests

Beta-binomial model The behavior of Huynh estimates of kappa (which are based on the beta-binomial model) did not appear different from their behavior in Study I, where distributions were clearly of the beta family. Huynh and Subkoviak estimates were at their highest observed values when the standards were at or near the distribution modes, regardless of the specific shape of the distributions.

Summary Kappa estimates were lower than rho estimates for the same test lengths, distribution shapes and standards. This was expected, kappa's regard for the chance level accounts for this difference. Kappa coefficients observed would not be considered reasonably high in terms of expected values for reliability coefficients.

Test length affected kappa estimates: increased test length increased the coefficients in every case. Both Huynh and Subkoviak estimates were at their highest observed values when the distribution modes and standards converged. This is the opposite of rho estimates, which were at their lowest observed values with this condition.

Comparison of Studies I and II

With simulated (Study I) and actual (Study II) data, both Huynh and Subkoviak estimates of rho were adequately high in terms of expected values for reliability coefficients.

The lowest observed rho coefficients were .67 and .74 for the Huynh and Subkoviak procedures, respectively. The two procedures yielded rho coefficients which differed very little from one another at comparable test lengths, distribution shapes and standards: the largest difference was .06.

Kappa estimates, on the other hand, differed considerably by estimation procedure in both studies. Subkoviak estimates for subtests were higher than comparable Huynh estimates by at least .03 in 47 (of 54) cases. Kappa coefficients were also consistently lower than rho coefficients, ranging as low as .107.

Test length affected all coefficients in a similar and unsurprising manner: as number of items increased, coefficients also increased. This was seen in every case without exception.

Distribution shapes had an overall impact on the levels of both rho and kappa estimates: the degree of skewness of a distribution affected the general level of estimates. In Study I, the normal distribution yielded the highest rho coefficients, followed (in order) by the slightly and highly skewed distributions. In Study II, the math test yielded the highest rho coefficients, followed (in order) by the slightly skewed geography test and more skewed biology test.

This phenomenon was due to the convergence of standards and modes in normal distributions and increasing divergence as distribution became more skewed. Convergence of standards and modes yields highest rho values.

An opposite, and less dramatic, effect of skewness was seen with the kappa estimates. In Study I, the normal distribution yielded the lowest coefficients, followed (in ascending order) by the slightly and highly skewed distributions. In Study II, the least-skewed math test yielded the lowest Huynh estimates, but not the lowest Subkoviak estimates. Additionally, the geography and biology tests did not follow the Study I pattern with either Huynh or Subkoviak kappas: the less skewed geography test yielded higher coefficients than the biology test. It should be noted, however, that the difference among kappa estimates for the three tests in Study II was only .07 for the Huynh estimates and .03 for the Subkoviak estimates.

Both rho and kappa estimates responded to the relationship of the distribution shape and the standard. Rho estimates were at their lowest observed value when the standard neared the distribution mode; kappa estimates were at their highest observed value in this condition. This was seen consistently with all simulated and actual data, regardless of test length and degree of skewness of distribution.

Studies I and II yielded no apparent differences in rho

or kappa behavior other than those accounted for by differences in distribution shape. Although the beta-binomial model, upon which Huynh estimates are based, apparently did not fit data in Study II, whether this was due to the goodness-of-fit procedure used (see Appendices A and B) or the robust nature of the model is unclear.

DISCUSSION

The current studies investigated the behavior of rho and kappa coefficients with simulated and actual tests of 25 or more items, various distribution shapes, and standards of 70%, 80% and 90% items correct. Huynh (1976) and Subkoviak (1976) estimates of rho and kappa for one-form administration were calculated.

All rho estimates were acceptably high, ranging from .67 to .99. Huynh and Subkoviak estimates of rho were of comparable magnitude in all cases. Kappa estimates were consistently lower, ranging from .10 to .77. The difference between rho and kappa levels was due to the consideration given to chance by the kappa coefficient. Huynh and Subkoviak estimations of kappa were more divergent than rho estimates, with Subkoviak coefficients being higher.

With both rho and kappa estimates, test length affected magnitudes of coefficients: increasing test length increased estimates without exception.

Magnitudes of coefficients were also affected by the skewness of the distribution shapes. Less skewed distributions (simulated and actual) yielded the highest rho coefficients and magnitudes decreased as skewness increased. With kappa estimates, the opposite was seen: the less skewed distributions yielded the lowest coefficients. However, for

kappa estimates, magnitudes did not change in a consistent direction with increases in skewness.

The effect of distribution shape on rho and kappa was a function of the proximity of distribution modes and standards. Rho was at its lowest value when the standard was near the distribution mode. In the current studies, with standards of 70%, 80% and 90% of items correct, the mode and standards converged only in the highly skewed distributions, resulting in lower overall estimates for these distributions. The opposite behavior occurred with kappa coefficients: kappa was at its highest magnitude when the standard and mode converged. Thus, the least skewed distributions yielded the lowest coefficients.

The behavior of rho and kappa in the current studies followed mathematical patterns and was thus, to a certain extent, predictable. Three phenomena, attributable to the nature of rho and kappa statistics, were apparent. First, values of rho were greater than values of kappa, due to the inclusion of the chance level in calculation of kappa. Second, longer tests yielded higher coefficients. Because scores were more spread in longer tests, lower proportions of students scored near the standards, yielding less inconsistency. Third, as discussed at length elsewhere, rho values were at their lowest when standards and distribution

modes converged.

No differences were seen between simulated and actual data, other than those attributable to differing distribution shapes.

Current and Previous Research

Findings in the current studies regarding the effect of the standard and distribution shape were in agreement with an earlier study with shorter tests and differing distribution shapes (Marshall & Serlin, 1979).

In both current and prior research, Huynh and Subkoviak procedures resulted in rho coefficients of similar magnitude (Subkoviak, 1978). No previous research, however, has compared kappa estimates. Huynh and Subkoviak estimates of kappa were not of similar magnitude in the current studies: Subkoviak estimates were consistently higher than Huynh estimates.

A possible explanation for the disparity of Huynh and Subkoviak kappa estimates lies in the method for calculating students' domain scores in the Subkoviak procedure. Subkoviak offers two methods for estimating domain scores: regression estimates and proportions of correct responses on the first form (p-values). In the current studies, p-values were used to compute domain scores. With large

sample tests (districtwide or statewide testing), assumptions of the regression model, namely, homogeneous subjects (and, perhaps, equal item difficulty), are unlikely to be met, mandating the use of p-values for domain estimates in these situations. Earlier research with Subkoviak's rho coefficient had concerned itself with smaller sample sizes (e.g., classroom samples) and mastery tests, wherein these assumptions are more likely to be met.

It should be unsurprising that the Subkoviak estimates based on the regression procedure yield estimates of rho and kappa of similar magnitude to Huynh estimates. The Huynh procedure (in assuming a beta-binomial model) implies that there is a linear regression of observed on true (hypothetical) scores. Both procedures, thus, restrict the range of scores and eliminate the fluctuations often seen in administered tests. (Preliminary findings in current research support the similarity of Huynh and Subkoviak kappa, as well as rho estimates when using Subkoviak's regression process.)

The disparity between Huynh and Subkoviak coefficients with the use of p-values in Subkoviak's calculations is less apparent for rho than kappa. This is due to the impact of the p-value twice in the calculation of kappa, in estimating both rho and the chance level.

In the current study using actual data, the apparent lack of fit of the beta-binomial model to data had no effect on the behavior of Huynh estimates. Problems regarding goodness-of-fit of the beta-binomial model notwithstanding (see Appendix A), current research supports the robustness of the model and, thus, the use of Huynh estimates. Previous research has, however, shown that in some conditions (e.g., extreme bimodality), the Huynh estimates do not function as expected (Marshall & Serlin, 1979).

Implications for Practitioners

A number of practical questions regarding the use of rho and kappa coefficients are likely to be asked by the practitioner.

(1) When is the use of rho and kappa appropriate? Before considering their use, the practitioner must be clear that the question being addressed concerns the consistency of categorization of students along a continuum, rather than the consistency of the degree to which an attribute is displayed. The latter is an issue of criterion-referenced score reliability (see Brennan and Kane, 1977) rather than reliability of mastery classifications, and is not addressed by rho or kappa.

(2) Which is better, one- or two-form estimates?

Two-form estimates are preferable in that they involve no bias and are easily calculable. They are advocated when two forms are, indeed, available and administered to examinees. When two forms are not available or not administered to the same examinees, the choice of one-form estimates, though psychometrically sound, involves loss of information regarding the status of those examinees who are classified inconsistently. One form estimates supply no information regarding the relative proportions of false masters and false nonmasters. While false masters may be of more consequence in mastery testing, false nonmasters are generally of more educational, legal and economic consequence in minimum competency testing. Although one form estimates given an indication of the proportion of students who may have been inconsistently classified, no indication of whether students are initially false-masters or false-nonmasters is available. This information would be desirable.

(3) Should a school calculate one-form estimates? In situations where only one form is available or administered to students, calculation of one-form estimates is contingent on several factors. First, the availability of computer programs for either Huynh or Subkoviak programs is limited, and both necessitate fairly sophisticated computers. Furthermore, the Subkoviak program, as currently available, does not

include estimation of kappa and must be so modified.

For schools with the necessary computer facilities, calculation of rho or kappa coefficients from one form tests provides valuable information about a test which is not available from more commonly found reliability coefficients. Although, in many instances, after a mastery or competency test is administered by a school district, the initial concern is the number of nonmasters (requiring remediation, retesting, etc.), knowledge of the degree of consistency in classification rendered by the test is valuable as an indicator of the degree of confidence that should be placed in categorizing examinees.

(4) Which are preferable: Huynh or Subkoviak estimates? Several factors must be considered in choosing estimation methods. The first, availability of computer resources, was discussed above. Second, the expected distribution shapes must be considered. In rare situations where distributions are not unimodal, use of the Huynh procedure is questionable in that it assumes a beta-binomial distribution. In the more usual situation, however, when scores are distributed unimodally, either estimation procedures is psychometrically acceptable. Third, the degree of bias (degree of deviation from two-score estimates) must be considered. Research with short tests has shown Huynh estimates to be stable and conservative, while Subkoviak estimates are less stable, both

under- and overestimating two-form estimates. While research has yet to uphold this finding with longer tests of varying distribution shapes, Huynh estimates may be the most cautious choice at this time.

(5) Shall I report rho or kappa? Rho, consistency of classification for any reason, and kappa, consistency of classification beyond chance, measure very different aspects of consistency. Whether one includes or excludes the role of chance is the major determinant of the choice between rho and kappa, whether they are based on one- or two-form coefficients.

Reporting of kappa coefficients is not without question when the standard is predetermined: chance level is figured by the proportion of examinees who are classified as masters on a particular test administration. This determination of chance level by the specific population is questionable (see Livingston & Wingersky, 1979), but in situations where populations vary greatly from testing to testing, the chance level and thus kappa may vary greatly between administrations of the same test.

Furthermore, the interpretation of kappa is problematic in that no acceptable levels are apparent. Whereas, the rho coefficient is on the more common 0 to +1 scale, the -1 to +1 scale of kappa is difficult to interpret.

(6) How do I use this information to evaluate and improve tests? It is very important that the practitioner understand the effect of the interaction between distribution shape and placement of the standard on both rho and kappa estimates. This effect is seen with both one- and two-form estimates.

Rho will be at its minimum observed value, and kappa at its maximum when the mode and standard converge. It is possible that a practitioner unfamiliar with this psychometric occurrence, will interpret a relatively low rho coefficient in this circumstance as an indication of problems with the test, and proceed to change the test or diminish the psychometric validity of the test. In a similar manner, the unaware practitioner may be more positive than warranted about a relatively high kappa estimate which results from this distribution - standard interaction. Such a judgment error is compounded by the lack of a clear range of acceptable values for kappa estimates.

In the development of a criterion-referenced test a major concern is the establishment of the appropriate difficulty level; in many cases, a difficulty level is considered appropriate if most (e.g., 70-80%) of students score above the standard. While necessary for the validity of the test, such an appropriate difficulty level creates a

paradoxical situation when rho and kappa are calculated. If a criterion-referenced test has a difficulty level allowing most students to pass, the distribution mode will likely be near the standard. As discussed above, rho will be at its minimum value and kappa at its maximum value in this situation. The practitioner is faced with a paradox in evaluating the goodness of the test: when the difficulty level is appropriate, rho and kappa are necessarily effected.

The varying magnitudes of estimates due to distribution shape and standard interaction must also be considered in the interpretation of long-term trends or comparisons between test scores of different populations. That is, in looking at the coefficients calculated from the same test at different administrations or from different populations, the practitioner may see very dissimilar values of coefficients. The overall worthiness of the test in consistently classifying examinees should be judged, when possible, on several calculations of rho and kappa rather than on one calculation with one population.

Implications for Future Research

A number of directions for future research are immediately apparent:

(1) Research similar to the present (test lengths of 25

or more, large sample sizes) with coefficients based on actual two-form tests is needed. Comparisons among two-form, Huynh and Subkoviak (both regression and p-value approach) estimates under these circumstances would provide an evaluation of the degree of bias (i.e., deviation from the two-form estimates) with longer tests and various commonly found distributions.

(2) Research with tests developed and used as criterion-referenced tests is also needed. The relationship between the standard and distribution mode on coefficients is a critical factor in estimates, particularly with extremely skewed distributions. Such extreme shapes are likely to occur with minimum competency tests, as items necessarily reflect instructional materials to which examinees have been exposed and it is expected that most of the students will be above the standard. The impact of the interaction between shape and standard when the standard is below the mode in an extremely skewed test is not fully known.

(3) The calculation of chance in the calculation of kappa is problematic when the standard is predetermined and applied to all examinees (see Livingston & Wingersky, 1979). As discussed above, changes in distribution shape greatly effect the calculation of chance, and thus, kappa. Perhaps

the use of a chance level which is based on several administrations of the same test or with several groups or examinees would provide an alternate procedure. This would lessen the common sample-to-sample fluctuations in calculation of chance and perhaps provide a clearer picture of kappa levels not tied to a specific population or test administration. While chance is, indeed, still based on distribution shape(s), the soundness and usefulness of such an approach should be explored.

(4) The relationship between classification consistency and criterion-referenced score consistency may offer needed information and possibly means for easier estimation of rho and kappa. What psychometric and practical relationship may there be between squared error loss (see Brennan & Kane, 1977) and rho and/or kappa? Can easily calculable statistics such as the standard deviation or the standard eror be adapted to provide information related to classification consistency? Are there ways to make estimates of rho and kappa that do not require the sophisticated computer programs used herein?

(5) Although the current Marshall & Serlin (1979) estimation procedure (wherein a hypothetical double-length test based on scores on one form is created and then split in half to create two tests) has received little research

due to its complexity and unavailability of computer programs, the merits of a split-half approach should be investigated. Such an approach may provide a method of calculating rho and kappa without the need to create a hypothetical second form, thus, avoiding both attendant assumptions and making calculation easier. For test of 50 or more items from the same domain and of approximately equal difficulty (nor an unlikely circumstance with minimum competency tests), a comparison between classification based on the two halves of the test may provide an adequate estimate of rho and kappa. In cases where one test samples from several domains, a splitting of items within each domain to create the two halves may be both practical and workable. Such an approach would be particularly valuable to schools with less sophisticated computer equipment.

REFERENCES

Aherns, J. H. & Dieter, V. Computer methods for sampling from gamma, beta, poisson, and binomial distributions. Computing, 1974, 12, 223-246.

Algina, J., & Noe, M. J. A study of the accuracy of Subkoviak's single administration estimate of the coefficient of agreement using true score estimates. Journal of Educational Measurement, 1978, 15, 101-110.

Beckham, J. Legal implications of minimum competency testing. Bloomington, Indiana: Phi Delta Kappa Education Foundation, 1980.

Berk, R. A. A consumer's guide to criterion-referenced test reliability. Journal of Educational Psychology, 1980, 17, 323-349.

Bloom, B. S. Mastery learning. In J. H. Block (Ed.), Mastery learning: Theory and practice. New York: Holt, Rinehart, & Winston, 1971.

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

Brown, F. G. Principles of educational and psychological testing (erd Ed.). New York: Holt, Rinehart & Winston, 1983.

Brown, F. G. Guidelines for test use: A commentary on the Standards for Educational and Psychological Tests. National Council on Measurement in Education, 1980.

Carver, R. P. Special problems in measuring change with psychometric devices. In American Institute for Research (Ed.) Evaluative Research: Strategies and methods. Pittsburgh, 1970, 48-63.

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 or partial credit. Psychological Bulletin, 1968, 70, 213-220.

DuBois, P. H. A history of psychological testing. Boston: Allyn and Bacon, Inc., 1970.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R. & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. Gagne (Ed.) Psychological principles in systems development. New York: Holt & Winston, 1962, 419-476.

Glaser, R. & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.) Educational measurement (2nd Ed.) Washington, D.C.: American Council on Education, 1971, 625-670.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Gross, A. L. & Shulman, V. The applicability of the beta-binomial model for criterion-referenced testing. Journal of Educational Measurement, 1980, 17, 195-202.

Hambleton, R. K. & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambleton, R. K., Mills, C. N. & Simon, R. Determining the lengths for criterion-referenced tests. Journal of Educational Measurement, 1983, 20, 27-38.

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.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Haney, W. Validity, vaudeville and values: A short history of social concerns over standardized testing. American Psychologist, 1981, 36, 1021-1034.

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

Herman, J. & Yeh, J. Test use: A review of the issues. In E. Baker & E. Quellmalz (Eds.) Educational testing and evaluation: Design, analysis and policy. Beverly Hills: Sage Publications, 1980, 219-228.

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

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. Psychometrika, 1977, 42, 601-608.

Huynh, H. Statistical inference for two reliability indices in mastery testing based on the beat-binomial model. Journal of Educational Statistics, 1979, 4, 231-246.

Huynh, H. Assessing efficiency of decisions in mastery testing. Journal of Educational Statistics, 1982, 7, 47-63.

Huynh, H. & Saunders, J. C. Accuracy of two procedures for estimating reliability of mastery tests. Journal of Educational Measurement, 1979, 4, 110-123.

Huynh, H. & Saunders, J. C. Solutions for some technical problems in domain-referenced mastery tests. National Institute of Education, Final Report, 1980.

Karabinus, R. A. & Ary, D. E. Flowchart of the process of selecting appropriate test reliability estimates. Paper presented at the Second Annual Symposium on Education, Chicago, May 1978.

Keats, J. A. Some generalizations of a theoretical distribution of mental test scores. Psychometrika, 1964, 29, 215-231.

Lerner, B. The minimum competency testing movement: Social, scientific and legal implications. American Psychologist, 1981, 36, 1057-1066.

Linn, R. L. Issues of validity for measurement in competency-based programs. In M. A. Bunda & J. R. Saunders (Eds.) Practices and problems in competency-based education. Washington, D.C.: National Council on Measurement in Education, 1979.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurment, 1972, 9, 13-26.

Livingston, S. A. & Wingersky, M. S. Assessing the reliability of tests used in making pass/fail decisions. Journal of Educational Measurement, 1979, 16, 247-260.

Livingston, S. A. & Zieky, M. J. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, N.J.: Educational Testing Service, 1982.

Lord, F. M. A strong true-score theory with applications. Psychometrika, 1965, 30, 239-270.

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

Mager, R. F. Goal analysis. Belmont, California: Fearon, Inc., 1972.

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.

Marshall, J. L. & Serlin, R. C. Characteristics of four mastery test reliability indices: Influence of distribution shape and cutting score. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1979.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan, 1974, 311-397.

Mitchell, S. K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 1979, 86, 2, 376-390.

Nitko, A. J. Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 1980, 50, 461-485.

Nunnally, J. C. Psychometric theory (2nd Ed.). New York: McGraw-Hill, 1978.

Peng, C. J., & Subkoviak, M. J. A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, 1980, 17, 359-368.

Popham, W. J. Educational evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Popham, W. J., & Baker, E. L. Establishing instructional goals. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

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

Reschly, D. J. Psychological testing for educational classification and placement. American Psychologist, 1981, 36, 1021-1034.

Resnick, D. Minimum competency testing historically considered. In D. Berliner (Ed.), Review of Research in Education, 1980, 8, 1-29.

Shepard, L. Technical issues in minimum competency testing. In D. Berliner (Ed.), Review of Research in Education, 1980, 8, 30-82.

Skakun, E. N., & Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

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

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

Subkoviak, M. J. Decision-consistency approaches. In R. A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980, 129=185.

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

Tables of the binomial probability distribution. Washington, D.C.: United States Government Printing Office, 1949.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison-Wesley Publishing Company, 1973.

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, 6, 3-32.

Zieky, M. L. & Livingston, S. A. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, New Jersey: Educational Testing Service, 1977.

ACKNOWLEDGMENTS

I wish to acknowledge the guidance and support of my advisor, Dr. Fred Brown, and committee members, Drs. Tom Andre, Mary Huba, Don Schuster, Bob Strahan and Rex Thomas. I also wish to thank Dr. Morris Wilson and Bill Schoenenberger, of the Des Moines Community Schools, for provision of data, computer assistance and encouragement in this research.

Mostly, I thank my parents, who made this endeavor probable, and my husband, Frank Seiler, who through his continuous patience and warmth, made it possible.

APPENDIX A: BETA-BINOMIAL ASSUMPTIONS

Two questions are addressed below: (1) What is a beta-binomial distribution?; and, (2) How can we determine if a distribution is a beta distribution?

The beta-binomial model assumes that estimates of ability parameters (e.g., test scores) are distributed within the population in a specified pattern. This pattern can assume a wide variety of shapes: normal, skewed, rectangular, U-shaped. All of the distributions in this family are unimodal (or U-shaped) and quasisymmetrical about the mode.

One mathematical formula defines the density function of all distributions in this family (see Wilcox, 1981). Alpha and beta are constants (parameters) used in this formula to determine (along with n, the number of items) the specific shape of each beta distribution. Alpha and beta have a mathematical correspondence to the first and second moments (the mean and variance) of a distribution.

Huynh uses the beta-binomial model in the generation of the hypothetical (second form) test scores. Use of the model is valuable, in that with all beta distributions, the regression of the true scores on the observed scores is linear

(Lord & Novick, 1968). In other words, the relationship between scores from the administered test and Huynh's hypothetical scores is linear.

A major concern with the use of this model lies with the lack of a statistical method for determining if the model provides a good fit to data (Wilcox, 1981). In cases where a distribution is unimodal or U-shaped "some member of the beta family should provide a good fit" (Gross & Shulman, 1980, p. 195). However, unimodal distributions have been reported that also apparntly were not beta distributions (Keats, 1964). In the case where a distribution of test scores is not unimodal, one assumes the model is not a good approximation to this data (Gross & Shulman, 1980).

Although there is no accepted statistical procedure to test the goodness of fit of the beta model, there are several methods of getting an indication of the goodness-of-fit. Three are reported briefly below.

Original work by Keats (1964) was extended by Wilcox (1981) in comparing an observed frequency distribution with the distribution predicted by the beta-binomial model. A chi-square statistic was used to compare the observed and expected frequencies.

Gross & Shulman (1980) made three predictions regarding expected frequencies, reliability coefficients, and validity

coefficients of the model given the parameters from a set of data. A descriptive comparison was then made between expected and actual data. The authors noted that "two theoretical predictions are in close agreement with the observed results" and that (results) "would be considered adequate by applied investigators" (p. 200). The prediction regarding validity coefficients was in less agreement with test data, though it may not have been "simply due to the inadequacy of the model" (p. 201).

Huynh and Saunders (1979) calculated maximum discrepancies between observed frequencies of several sets of data and those expected from the model. These discrepancies were then subjected to the Kolomogorov-Smirnoff test and the significance level obtained. Conclusions were that several sets of data "follow closely the beta-binomial model" and others "reveal substantial departures" (p. 114).

In terms of the subsequent behavior of rho estimates, Huynh and Saunders found no difference in size, direction or degree of error between those data sets that followed and those that departed from the model. Other research (Wilcox, 1981) also supported the robust nature of the model: in cases were the model apparently did not fit data, no difference in behavior of Huynh's estimates was seen.

APPENDIX B: GOODNESS-OF-FIT TEST FOR THE BETA-BINOMIAL MODEL

The observed frequencies for each of the three full-length tests and nine subtests were individually compared with expected frequencies from a beta distribution derived using the same alpha and beta parameters as the observed distribution. The Aherns and Dieter (1974) algorithm to generate random numbers from a beta distribution was used to generate nonzero scores for ten times the number of examinees on each test. This ten-fold sample was used to provide more accurate estimation. The frequencies generated were then divided by ten for comparison with observed frequencies.

A chi-square statistic was calculated for each test distribution, comparing the observed and expected (i.e., generated) frequences. Frequencies at the tails of the distributions were summed to eliminate cells with zeros. This summing of cells at the tails was done with all distributions, though more so with the highly skewed ones. Degrees of freedom were adjusted accordingly.

All chi-square statistics were significant at the $p<.01$ level, as shown in Table 9. In comparing observed and expected distributions, it was apparent that large deviations often appeared at the tails of the distributions. The small

cell sizes at the tails had a greater impact on the chi-square statistic relative to cells in the middle of the distributions. Thus, it may be that the finding of significance was, in at least some cases, due to the combination of a large sample size (reducing the impact of deviations in the middle of the distributions), and relatively large deviations between observed and expected frequencies at the distribution tails.

Appendix A discussed the controversial (and tentative) nature of measuring the goodness-of-fit of the beta-binomial model to data. Whether the current findings are due to these concerns, the use of the chi-square statistic with large sample sizes and large deviations at the tails, or simply because all observed distributions were, indeed, not of the beta family, is unclear.