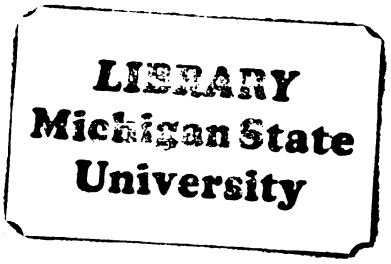
THESIS



This is to certify that the

dissertation entitled

TEST CHARACTERISTICS AND THE BIAS AND SAMPLING VARIABILITY

OF CRITERION-REFERENCED RELIABILITY COEFFICIENTS

presented by

Loraine Son

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Psychology

Major professor

Date 2/21/83

MSU is an Affirmative Action/Equal Opportunity Institution

0-12771

MSU
LIBRARIES

RETURNING MATERIALS:
Place in book drop to remove this checkout from your record. FINES will be charged if book is returned after the date stamped below.

TEST CHARACTERISTICS AND THE BIAS AND SAMPLING VARIABILITY OF CRITERION-REFERENCED RELIABILITY COEFFICIENTS

by

Loraine Son

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

1983

TES

The
the majo
coeffici
categori
upon whi
agreemer
differe
shapes,
binomia
the stu
tions c
fluctua
sizes.
a psyc
as tes
invest
were c

ABSTRACT

TEST CHARACTERISTICS AND THE BIAS AND SAMPLING VARIABILITY OF CRITERION-REFERENCED RELIABILITY COEFFICIENTS

By

Loraine Son

The present study examined the bias and sampling variability of the major single test administration criterion-referenced reliability coefficients given various test parameters. These coefficients were categorized by the type of loss function (squared error or threshold) upon which they were based and by whether they included a chance agreement correction. The extent of bias was studied as a function of different parallelism conditions (classic versus random), distribution shapes, and cut-off scores. Distributions not belonging to the beta-binomial family and the random parallelism condition were included in the study to examine the robustness of several coefficients to violations of their underlying assumptions. Each coefficient's sampling fluctuation was investigated for various test lengths and sample sizes. Data from the Michigan Educational Assessment Program and from a psychology mid-term exam were used to generate item domains as well as test scores, and were altered to reflect the various parameters investigated. For each cell of the design, population coefficients were computed from either randomly or classically parallel alternate

forms.

populati

single t

samples.

the coef

For

coeffic

the cla

paralle

not alw

approac

became

geneous

of the

general

kappa

beta-b

test

recomm

catego

forms. To determine the magnitude and direction of bias, the population values were compared to the mean of the corresponding single test administration sample estimates taken across many samples. The standard deviation of these sample estimates indicated the coefficient's sampling variability.

For distributions derived from homogeneous item domains, all the coefficients, except the kappa estimates, were robust to violation of the classic parallelism assumption. For the other distributions, the parallelism condition did affect the coefficients' biases, although not always in the expected direction. Generally, as the cut-off score approached a distribution's mean, the squared error coefficients became more biased for randomly parallel tests consisting of heterogeneous items. The cut-off score also significantly affected the bias of the threshold agreement coefficients. However, the results, generally, did not follow a pattern. The hypothesis that the p_o and kappa estimates would be more biased for distributions which were not beta-binomial was unsupported. Sampling variability decreased as the test length and sample size increased. Based on these results, recommendations were made about which coefficient to use within each category given various test conditions.

I

person

whose

his sa

during

stage

Dr. W

advic

under

memb

am i

comp

thos

work

supp

but

hum

con

the

ACKNOWLEDGEMENTS

I wish to express my most sincere and warmest thanks to my chairperson, Dr. Neal Schmitt, who introduced me to this area of study and whose suggestions guided me through the most difficult times. Without his sage advice as well as his calm, rational, and nurturant manner during these periods, the project may never have reached its final stages. His patience and support have been invaluable.

I also wish to express my appreciation to Dr. Raymond Frankmann, Dr. William Mehrens, and Dr. Frederic Wickert for their timely, needed advice and their willingness to respond with constructive feedback under the pressure of a short time frame. Apart from my committee members, other individuals have contributed notably to this project; I am indebted to Bryan Coyle for his kindness in offering to apply his computer expertise when needed, and to Kathy Sigafoose as well as those who assisted her, Kathy Cooper and Janet Larrimore, for the hard work and time they devoted to typing and preparing this manuscript.

Finally, I wish to thank my parents for not only providing love, support, and patience through the more trying times of this project, but also for their understanding, kindness, devotion, and sense of humor which have guided me throughout my life. For these priceless contributions and with much love and respect, I dedicate this work to them.

TABLE OF CONTENTS

Table

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

LIST OF TABLES

13. Me

14. Mea

15. Mea

16. Di

17. Rec

Figure

1. Jo

2. Ma

3. Ad

4. Sk

5. J-

6. Bi

7. No

8. Fo

LIST OF FIGURES

to wh

occas

items

1976)

relia

tatio

appli

becom

ment,

indiv

Husek

assoc

(e.g.,

alpha

remai

do not

examp

making

indiv

addres

gists

TEST CHARACTERISTICS AND THE BIAS AND SAMPLING VARIABILITY OF CRITERION-REFERENCED RELIABILITY COEFFICIENTS

Reliability denotes the consistency of measurement or the extent to which scores are reproducible over repeated testings on different occasions, or over different sets of parallel or randomly parallel items, and/or under other small variations in conditions (Anastasi, 1976). Although it is frequently stated that a test is reliable, reliability actually refers to the consistency of the score interpretation obtained from the test, not to the test, in and of itself.

In industrial-organizational psychology as well as in other applied sciences, norm-referenced interpretations of test scores have become the sine qua non of measurement. In norm-referenced measurement, an individual's score is given meaning by determining the individual's relative standing within a normative group (Popham & Husek, 1969). Quite appropriately, those reliability coefficients associated with classical test theory and norm-referenced measurement (e.g., correlation between two test administrations, coefficient alpha) indicate the extent to which examinees' relative standings remain consistent. However, norm-referenced interpretations of data do not satisfy all the measurement needs of psychologists. For example, in many situations, scores ultimately serve as a basis for making dichotomous decisions (e.g., accept/reject) or placing individuals into groups (e.g., successful/unsuccessful). In order to address these and other measurement needs, many educational psychologists and measurement experts have turned to criterion-referenced

measure

indicat

data ar

No

Glaser

As prev

relevan

perform

these s

scores,

On

levels

ment co

Klaus,

demonst

are com

testing

percent

anchor

(Hamble

made as

particu

perform

classif

his/her

measurement and, in so doing, have had to create coefficients indicating the extent to which criterion-referenced interpretations of data are reliable.

Criterion-Referenced Measurement and Tests

Norm- and criterion-referenced measurements were distinguished by Glaser (1963) on the basis of the standard used to interpret scores. As previously noted, the former uses the test scores of members of a relevant group as the standard for judging an individual's performance. Consequently, the mean serves as the anchor point of these scales and raw scores are typically converted into standard scores, percentiles, stanines, or ranks (Eignor & Hambleton, 1979).

On the other hand, criterion-referenced measurement uses defined levels of criterion behavior along an achievement, skill, or attainment continuum as the performance standard (Glaser, 1963; Glaser & Klaus, 1962). More specifically, the behaviors required to demonstrate competence at each proficiency level are identified and are compared to the behaviors exhibited by an individual on the testing instrument. A typical criterion-referenced measure is the percentage of items answered correctly. This type of scale has two anchor points, one at each end of the scale, i.e., 0% and 100% (Hambleton & Eignor, 1979). In most circumstances, some evaluation is made as to whether or not an individual's score indicates mastery of a particular skill, objective, etc.; a minimally acceptable level of performance, cut-off score, is established and an individual is classified as either a master or non-master according to whether his/her score is above or below this predetermined level of competence

(Buck,

several

priate

referer

"indepe

p. 520;

employ

based u

Si

appropr

The star

various

Anderson

criterio

necessit

individua

objective

skills a

proficien

tions hav

organizat

placement

multiple c

1962; Gold

been the d

others' per

individual

(Buck, 1975; Hambleton & Novick, 1973). In other applications, several cut-off scores may be used to divide the examinees into appropriate groups. Contrary to norm-referenced measures, criterion-referenced scores indicate what an individual can and cannot do "independent of reference to the performance of others" (Glaser, 1963, p. 520; Glaser & Klaus, 1962). In short, norm-referenced measures employ a relative standard, while criterion-referenced measures are based upon an absolute standard (Glaser, 1963).

Situations where criterion-referenced measurement would be more appropriate than norm-referenced measurement are easily discernible. The standards used indicate that both are appropriate for making various decisions about individuals (Popham & Husek, 1969; Wardrop, Anderson, Hively, Hastings, Anderson, & Muller, 1982). In education, criterion-referenced measurement gained prominence partly due to the necessity of diagnosing student needs and assessing performance in individualized instructional programs (Mehrens & Ebel, 1979). The objective of measurement in these instances was to determine what skills a student possessed or to simply assess whether a student was proficient in a particular area. Similar types of score interpretations have long existed in various aspects of industrial-organizational psychology such as performance appraisal, job placement, training performance, and personnel selection via the multiple cut-off and the multiple hurdle techniques (Glaser & Klaus, 1962; Goldstein, 1974). In all these areas, a frequent concern has been the determination of an individual's performance independent of others' performance, and decisions have been typically made about the individual's mastery of an objective. Such score interpretations have

been par

al., 19

Cri

decision

referenc

Popham &

the with

is desir

(Popham

this cas

post-tes

As

measurem

drawing

(Glaser,

tation a

built us

of the c

can do (

research

tests.

Var

offered.

criterio

been particularly prevalent within a free quota system (Wardrop et al., 1982).

Criterion-referenced measurement is also appropriate for making decisions about treatments (e.g., training programs), while norm-referenced measurement is not as suitable (Mehrens & Ebel, 1979; Popham & Husek, 1969). The latter technique is designed to increase the within group variance, while having a small within group variance is desirable when evaluating the effects of different treatments (Popham & Husek, 1969). A typical criterion-referenced measure in this case is the proportion of individuals achieving mastery on a post-test.

As can be seen in the above examples, criterion-referenced measurement is not concerned with rank-ordering individuals, but with drawing conclusions about an individual's behavioral repertoire (Glaser, 1963). The psychometric implications of this score interpretation are far-reaching. Several experts have suggested that tests built using classical methods do not provide adequate representation of the content needed to make generalizations about what an individual can do (Glaser & Nitko, 1971; Hambleton & Novick, 1973). In response, researchers have focused on the development of criterion-referenced tests.

Various definitions of criterion-referenced tests have been offered. Ivens (1970) proposed the following general definition: "A criterion-referenced test is one composed of items keyed to a set of

behavioral objectives" (p. 2). In comparison, a very specific and restrictive definition was offered by Harris and Stewart (1971):

> A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed (p. 1).

Similarly, Glaser and Nitko (1971) advanced the following definition: "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Expanding upon this definition, Glaser and Nitko (1971) stated:

> Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain, and such measurements are referenced directly to this domain for each individual measured (p. 653).

These definitions of criterion-referenced tests are sufficiently different that a particular test could be classified as either norm- or criterion-referenced, or could contain characteristics of each depending upon the definition adopted (Hambleton & Novick, 1973). However, all these definitions imply that criterion-referenced tests are constructed by and dependent upon the existence of a well-specified content domain as well as procedures for generating samples of items from this domain (Hambleton & Novick, 1973).

Some measurement experts questioned the accuracy and relevance of distinguishing between norm- and criterion-referenced tests (Brennan, 1979; Hambleton & Novick, 1973; Mehrens & Ebel, 1979). On the other hand, Hambleton and Eignor (1979) contended that a distinction should be made since a methodology now exists for constructing the latter

tests. Mehrens and Ebel (1979) observed that any test, whether it be criterion- or norm-referenced, represents a specified content domain. Moreover, a criterion-referenced test can be used to make norm-referenced measurements and, conversely, criterion-referenced measurements can be derived from norm-referenced tests, although neither of these usages may be very satisfactory (Hambleton & Novick, 1973). Given these facts, the primary distinction appears to be between norm- and criterion-referenced measurement (i.e., interpretation) rather than between different types of tests (Brennan, 1979; Ebel, 1971; Hambleton & Novick, 1973; Mehrens & Ebel, 1979). Of course, choosing a particular type of measurement prior to test construction has different implications for the method used to determine the items to be included on a test. However, Brennan (1979) proposed that different methods of test construction and item analysis produce changes in the definition of the item domain relevant for each measurement type rather than effect changes in the measurements themselves. The point is that scores can be interpreted using either standard for most tests. The legitimacy of such an interpretation is a different issue and depends upon the manner used to construct the tests as well as how restrictive a definition one adopts for a criterion-referenced test (Mehrens & Ebel, 1979; Wardrop et al., 1982).

In preparing test items for a criterion-referenced score interpretation, the overriding interest is how well the item samples the content domain or criterion behavior (Wardrop et al., 1982). The reason for such concern is the need for generalizing from specific test item responses to the whole domain of behaviors in order that inferences can be made about what skills the examinee possesses

(Hambleton & Eignor, 1979). Although test development for norm-referenced measurement is frequently concerned with defining the domain of interest, criterion-referenced testing involves far more concern with this issue and with obtaining a representative sample of items from this domain (Hambleton & Novick, 1973; Wardrop et al., 1982). In short, the basic steps in constructing tests specifically for criterion-referenced measures are specifying the domain, writing items reflecting these specifications, and selecting items via a sampling procedure (random or stratified random sampling, representative sampling) which assures representativeness. Similarly, the primary approach for conducting an item analysis after test construction is to have content specialists judge whether each item appropriately measures some part of the content domain as well as whether the items adequately sample the domain (Buck, 1975; Hambleton & Eignor, 1979).

An objective of tests designed for norm-referenced measurement has been to maximize variability so that individuals can be reliably rank-ordered. Norm-referenced measures, such as standard scores, depend upon the existence of variability since they are derived by comparing an individual's scores to the scores of a relevant group. Variability is partly achieved by using classical test development methods to analyze items. The assumption underlying classical methods is that a measurement procedure should provide the most discrimination possible among individuals on a particular characteristic. Consequently, items are largely analyzed and chosen based on their statistical characteristics, e.g., discrimination index, difficulty level, and item-total correlation.

On the other hand, the need for a "criterion-referenced test" to produce variability has been the topic of some debate. Some theorists have contended that variability is irrelevant to criterion-referenced measurement since these scores derive their meaning through a direct comparison with the performance criterion (Millman & Popham, 1974; Popham & Husek, 1969). In addition, many applications (e.g., a post-training test) exist in which the goal may be to have every examinee in the sample achieve mastery and, in so doing, to actually restrict the test score range. In contrast, Woodson (1974a) has argued that a "criterion-referenced test" must produce variability or else it is not informative or useful. The premise for Woodson's argument was that a test should be analyzed and developed on observations representative of those within the range of interest and, as a result, should discriminate between different observations of the characteristic. Using this approach, one would include pre- and post-training test scores in the range of possible observations used to calibrate an instrument (Woodson, 1974a). No variance may exist within the pre-training test nor within the post-training test, but the test should discriminate between these two testing observations. In contrast, Millman and Popham (1974) contended that the population of observations for a test designed to elicit criterion-referenced measures is "a domain of items and the responses of a single individual to them" (p. 137). Furthermore, they stated that if items were chosen on the basis of their ability to discriminate between observations, the test would not contain a sample of items truly representative of the content domain. The major difference between these two positions clearly lies in defining the appropriate group to

be used for calibrating the scale (Woodson, 1974b). Proponents of both sides agree, however, that items should not be chosen so as to maximize test score variance (Woodson, 1974b). Therefore, the variability can be expected to be lower than for "norm-referenced tests". Moreover, in typical usage, the test score variance may be very limited or non-existent if a test is administered to a sample of examinees who have just completed an instructional program.

Reliability

The possible absence or dimunition of score variability and, more importantly, the type of score interpretation associated with criterion-referenced measurement make classical reliability estimates inappropriate for indexing the consistency of these measures. In classical test theory, the reliability coefficient equals the squared correlation between true scores and obtained scores, i.e., the ratio of true to obtained score variance. All of the practical formulations (e.g., correlation between classically or randomly parallel tests, coefficient alpha, split-half reliability) for estimating this ratio require the computation of a correlation coefficient whose size is largely a function of the amount of variability in the sample. As is well known, the more heterogeneous the sample, the higher the reliability coefficient. This fact is easily seen from the equation for reliability: $r_{11} = s_T^2 / s_t^2 = 1 - (s_e^2 / s_t^2)$ where r_{11} equals the reliability and s_T^2, s_e^2, and s_t^2 denote the true score variance, the error variance, and the total score variance, respectively. For any given test, the error variance remains the same from sample to sample, regardless of the size of the total variance, because the size of the

error only depends upon the test's inability to provide accurate measures of individual true scores (Magnusson, 1967). However, the total and true score variances increase when a more heterogeneous sample is given the test, resulting in a larger reliability coefficient. Conversely, when no true score variance exists, the reliability equals zero (unless $s_e^2=0$, in which case, reliability is undefined). Due to this dependence upon score variability, a test used for criterion-referenced measurement might be highly consistent in a test-retest sense, and yet the classical reliability estimates might deem it to be unreliable because almost everyone has received the same score. A criterion-referenced measure might even have a negative internal consistency index and still be a reliable measure (Popham & Husek, 1969). In short, classical reliability estimates provide an unjustified pessimistic view of the consistency of criterion-referenced measurement due to the former's dependence upon variability (Buck, 1975). High classical reliability estimates can be used to support a claim of consistency, but low estimates do not indicate a lack of reliability (Popham & Husek, 1969).

As noted previously, criterion-referenced measurement most commonly involves mastery assessment where one cut-off score is used as the performance standard. Therefore, reliability for this type of measurement should assess the dependability of the mastery decision. Clearly, classical reliability estimates are insensitive to this type of consistency. Since reliability is based on the relationships between true, observed, and error scores, this viewpoint can be presented more clearly by determining the impact of mastery score interpretation upon these variables and their relationships. Marshall

(1978) provided an excellent discussion in this area, and much of the following material was derived from his presentation.

In classical test theory, the relationship among obtained score ($\underline{X}$), true score ($\underline{T}$), and error ($\underline{E}$) is expressed by the well-known equation $\underline{X}=\underline{T}+\underline{E}$. The distributions of true and error scores are continuous, while $\underline{X}$ has a polytomous or many-valued discrete distribution (Marshall, 1978). Theoretically, obtained scores could have a continuous distribution, but measurement instruments do not provide the necessary discriminations (Marshall, 1978). The effect of mastery testing upon this basic equation can be easily seen if testing is viewed within a decision-theoretic framework (Hambleton & Novick, 1973). In mastery testing, one wants to decide whether an examinee's true performance level is above or below a threshold or cut-off score; mastery testing can be viewed as a classification problem (Hambleton & Novick, 1973). Therefore, the comparable equation for mastery testing is $\underline{D}=\underline{C}+\underline{M}$ where $\underline{D}$, $\underline{C}$, and $\underline{M}$ represent the obtained classification, the true classification, and the instance as well as the direction of misclassification, respectively (Marshall, 1978). This model differs from its classical test theory counterpart in that all the variables in the equation are discrete as well as dichotomous given the absolute value of the misclassification error (Marshall, 1978). Viewed in this way, mastery testing results in a Platonic true score model (Marshall, 1978).

Using a Platonic instead of a classical true score model for mastery testing has implications for the determination of reliability. First, according to Marshall (1978), statistics such as

a mean or a

model since

cannot be a

point is de

properties

interval e

fication e

Second, me

In classic

error. Ho

defined in

i.e., the

(Marshall

not be hi

issue in

examinee

a retest

Swaminat

reliabil

repeated

Platonic

reliabil

one depe

examinee

(Hambleto

reliabi

therefo

a mean or a variance are "theoretically not meaningful" for the former model since observed and true scores in the Platonic true score model cannot be attributed with more than ordinal properties (p. 4). (This point is debatable; one could argue that these scores have interval properties when only two mastery levels exist since there is one interval equal to itself.) The absolute value of the misclassification error can also be assumed to be ordinal (Marshall, 1978). Second, measurement error is defined differently for the two models. In classical test theory, one is concerned with the size of the error. However, in the Platonic true score model, error can only be defined in terms of the existence of misclassification, not its size, i.e., the examinee is either correctly or incorrectly classified (Marshall, 1978). Moreover, these two types of measurement error need not be highly correlated (Marshall, 1978). Given these facts, the issue in assessing reliability for mastery testing is whether an examinee is assigned to the same mastery state on parallel tests or on a retest (Hambleton & Eignor, 1979; Hambleton & Novick, 1973). Swaminathan, Hambleton and Algina (1974) defined mastery testing reliability as "the measure of agreement between the decisions made in repeated test administrations" (p. 264). Consequently, given the Platonic true score model, the appropriate loss function for reliability estimation is threshold loss, where loss is either zero or one depending upon whether the two testing procedures assign the examinee to the same or to different mastery states, respectively (Hambleton & Novick, 1973; Marshall, 1978). The correlational reliability estimates use a squared error loss function and are, therefore, inappropriate (Hambleton & Novick, 1973).

Some corr

ficients, do u

appropriate fo

(1978) examine

measure the sq

classificatio

states. The

theoretical a

ficient is on

argue that th

underlying va

is somewhat a

problem is th

correctly re

classificati

true negativ

negative cla

negative (Ma

Thi
wer
all
or
ma

One other p

respect to

cation. In

Contra

coefficient

dichotomize

Some correlational statistics, the phi and the tetrachoric coefficients, do use a threshold loss function and, therefore, might seem appropriate for assessing reliability in the Platonic model. Marshall (1978) examined the ability of these coefficients to accurately measure the squared correlation between the obtained and true classifications (i.e., classical reliability) given two mastery states. The phi coefficient was found to be deficient on both theoretical and statistical grounds. As is well known, the phi coefficient is only appropriate for true dichotomies. One can easily argue that the mastery/non-mastery dichotomy is artificial since the underlying variable is continuous and the setting of the cut-off score is somewhat arbitrary (Glass, 1978; Marshall, 1978). The statistical problem is that phi can be negative when a negative value does not correctly reflect the relationship between the true and obtained classifications. More specifically, if either the true positive or true negative classification is zero and the false positive and false negative classifications are non-zero, the phi coefficient will be negative (Marshall, 1978).

> This would mean, for instance that even though there were only a few true non-masters (5%, say), if they are all misclassified then phi is negative, even though 90% or more of the examinees are correctly classified as masters (Marshall, 1978, p. 7).

One other problem with phi occurs when no variability exists with respect to the true mastery status and/or the obtained classification. In this instance, phi is undefined.

Contrary to phi, the use of the tetrachoric correlation coefficient is appropriate when the two variables are artificially dichotomized. However, this coefficient assumes that the two

variables hav

score distrib

Since di

Marshall (197

coefficient.

stantial numbe

cannot be solv

data points si

the phi coeffi

previously dis

Since none

Marshall (1978

to the obtaine

ρ are the true

respectively.

frequencies sh

(A+2) (B+2) (M

problems with

one, regardless

Second, the ra

Finally, if th

equal zero, an

similar to tha

In conclusion,

to obtained mas

of the data eve

function.

variables have a bivariate normal distribution while mastery test score distributions are often bimodal (Marshall & Serlin, 1979).

Since dichotomous variables have ordinal data properties, Marshall (1978) also considered the Spearman rank order correlation coefficient. However, this statistic is inappropriate when a substantial number of tied ranks exist (Marshall, 1978). This problem cannot be solved by computing the Pearson $\underline{r}$ using the tied ranks as data points since the resultant formula is algebraically equivalent to the phi coefficient and, consequently, is subject to the problems previously discussed (Marshall, 1978).

Since none of the correlational approaches proved satisfactory, Marshall (1978) examined the ratio of the true classification variance to the obtained classification variance: $\pi(1-\pi) \,/\, \underline{p}(1-\underline{p})$ where π and $\underline{p}$ are the true and obtained proportions of mastery classification, respectively. This formula can also be expressed in terms of the cell frequencies shown in Figure 1, i.e., $\pi(1-\pi) \,/\, \underline{p}(1-\underline{p}) = (\underline{A}+\underline{B})\,(\underline{C}+\underline{D}) \,/\, (\underline{A}+\underline{C})\,(\underline{B}+\underline{D})$ (Marshall, 1978). Marshall also found several statistical problems with this formula. First, if $\underline{A}=\underline{D}$ or $\underline{B}=\underline{C}$, the ratio equals one, regardless of the frequencies contained in the other two cells. Second, the ratio can be greater than one if $.5 \leq \pi < \underline{p}$ or $\underline{p} < \pi \leq .5$. Finally, if the obtained score variance is zero, either $\underline{p}$ or $1-\underline{p}$ must equal zero, and the ratio will be undefined. The latter problem is similar to that of the typical correlational reliability estimates. In conclusion, none of the correlational indices nor the ratio of true to obtained mastery score variance adequately reflect the reliability of the data even though all these indices use a threshold loss function.

True

Figure 1.--

Not ever
model and, c
testing (Bre
Although mo
individuals
far an exam
of reliabil
classical
bility of
uses a squ
latter req

One o
of classic
ment. Doe
underlying
(1) E=X-T;
(3) ρ(T,E)
1968). (T
(1974) not
score equa

True Classification	Obtained +	Obtained -
+	A	B
-	C	D

Figure 1.--Joint Distribution of True and Obtained Classifications

Not everyone in the field agrees that the Platonic true score model and, consequently, threshold loss are appropriate for mastery testing (Brennan, 1979; Kane & Brennan; 1977; Livingston, 1972b). Although more will be said about this viewpoint at a later time, these individuals contend that the major question in mastery testing is how far an examinee's score is from the cut-off. This conceptualization of reliability uses a squared error loss function. Despite this fact, classical estimates are still inappropriate for evaluating the reliability of this criterion-referenced interpretation because the former uses a squared error loss function with respect to the mean while the latter requires squared error loss with respect to the cut-off score.

One other issue should be addressed in judging the applicability of classical reliability estimates to criterion-referenced measurement. Does criterion-referenced measurement satisfy the assumptions underlying classical test theory? Briefly, these assumptions are: (1) $\underline{E}=\underline{X}-\underline{T}$; (2) $\varepsilon(\underline{E})=0$ in every non-null subpopulation of individuals; (3) $\rho(\underline{T},\underline{E})=0$; (4) $\rho(\underline{E}_1,\underline{E}_2)=0$; and (5) $\rho(\underline{E}_1,\underline{T}_2)=0$ (Lord & Novick, 1968). (The subscripts 1 and 2 denote parallel tests.) Brennan (1974) noted that $\varepsilon(\underline{E})$ cannot be zero for the subpopulation with true score equal to the highest value nor for the subpopulation with true

score equal to zero. Furthermore, Marshall (1978) found that these assumptions did not fare very well under the Platonic true score model.

> Klein and Cleary (1967) have shown, among other things, that with the Platonic true-score model, the correlation of true and error scores is generally negative and is zero only under extraordinary circumstances, that the expected value of Platonic error is not likely to be zero, and that errors on parallel tests cannot be expected to have zero correlation (Marshall, 1978, p. 5).

To summarize, classical reliability estimates used in norm-referenced measurement are inappropriate for criterion-referenced measurement, in general, because of the former's dependence upon score variance. In addition, classical estimates are particularly unsatisfactory for mastery assessment for two other reasons: (1) they use an inappropriate loss function, squared error loss with respect to the mean, and; (2) the classical assumptions underlying their use may not be met if the Platonic true score model is accepted as the model for mastery measurement.

Several reliability coefficients have been developed for criterion-referenced measurement. Since mastery assessment is involved in the vast majority of cases, most of the coefficients have been proposed within this context. Basically, reliability formulations for mastery assessment can be divided into two types based upon whether they use a threshold loss function or a squared error loss function with respect to the cut-off score. Of course, the choice of a loss function depends upon one's definition of the purpose of mastery testing. To reiterate, advocates of threshold loss contend that mastery testing is a matter of classifying examinees into two or

possibly more mutually exclusive categories (Hambleton & Novick, 1973; Kane & Brennan, 1977). Proponents of the opposing view assert that the issue is the "degree to which the student has attained criterion performance", implying that the estimation of the distance between the examinee's score and the cut-off score is the major concern (Glaser, 1963, p. 519). Both types of coefficients are presented below along with studies evaluating their performance under a variety of test characteristics.

Reliability Formulations Based Upon Squared Error Loss

Livingston. Livingston (1972b) developed a general form of the typical reliability coefficient applicable to both criterion- and norm-referenced measurement. He adapted the classical test theory model by replacing the deviation of scores about the mean with the deviation of scores about the cut-off in computing all relevant statistics. For example, he defined a criterion-referenced correlation coefficient as a product-moment parameter based on moments about the cut-off. The classical test theory assumptions, the traditional definitions of true score and errors of measurement, and the well-known relationships among true, error, and observed scores remained intact in his formulation. Corresponding to the classical reliability definition, Livingston defined criterion-referenced reliability as the squared criterion-referenced correlation between observed and true scores. This definition in conjunction with the classical test theory assumptions resulted in the following formula:

$$\underline{K}^2(\underline{X},\underline{T}_{\underline{X}}) = \frac{(\rho^2(\underline{X},\underline{T}_{\underline{X}})\ (\sigma^2_{\underline{X}})) + (\mu_{\underline{X}}-\underline{C}_{\underline{X}})^2}{\sigma^2_{\underline{X}} + (\mu_{\underline{X}}-\underline{C}_{\underline{X}})^2} = \frac{\sigma^2_{\underline{T}} + (\mu_{\underline{X}}-\underline{C}_{\underline{X}})^2}{\sigma^2_{\underline{X}} + (\mu_{\underline{X}}-\underline{C}_{\underline{X}})^2}$$

where C_x is the cut-off score, $\rho^2(X,T_x)$ equals any norm-referenced reliability coefficient, σ_x^2 is the variance, μ_x is the mean, and σ_T^2 equals the true score variance (Livingston, 1972b).

Lovett (1977) defined $K^2(X,T_x)$ in analysis of variance terms. He assumed the data came from the responses of n individuals to k parallel measurements or items, resulting in nxk observations. The design of the ANOVA was a randomized complete block design without interaction. Given that the only major differences in using ANOVA to estimate the reliability of norm-referenced versus mastery scores are the degrees of freedom for both the total sum of squares and the sum of squares for people, and the substitution of C_x for the mean in all relevant statistics, Lovett (1977) defined the reliability of mastery measurement as:

$$\frac{E(MS_p)-E(MS_e)}{E(MS_p)}$$

where all sums of squares are expressed as deviations from C_x.

Brennan and Kane. Brennan and Kane's "index of dependability" was derived from generalizability theory rather than from classical test theory (Brennan & Kane, 1977a). Two major differences between these theories are: (1) classical test theory is built upon the assumption of classically parallel tests, while generalizability theory assumes random parallelism; and (2) classical test theory does not differentiate among various types of errors while generalizability theory does (Brennan, 1978; Cronbach, Gleser, Nanda, Rajaratnam, 1972). This differentiation is accomplished by constructing a general linear model of the data and using analysis of variance procedures to

derive a reliability coefficient. Since a complete understanding and derivation of Brennan and Kane's index requires a very lengthy discussion, only a brief outline of their analysis has been provided below.

Assuming that test data have been derived from a random sample of items from an infinite domain of items and a random sample of people from an infinite population, the following linear model represents the observed score of person "$\underline{p}$" on item "$\underline{i}$":

$$X_{\underline{pi}} = \mu+\mu_{\underline{p}}{}_{\sim}+\mu_{\underline{i}}{}_{\sim}+\mu_{\underline{pi}}{}_{\sim} \quad \text{where}$$

μ = grand mean in the population of persons and the domain of items

$\mu_{\underline{p}}{}_{\sim}$ = effect for person $\underline{p}$

$\mu_{\underline{i}}{}_{\sim}$ = effect for item $\underline{i}$

$\mu_{\underline{pi}}{}_{\sim}$ = effect for the interaction of $\underline{p}$ and $\underline{i}$ plus experimental error (Brennan, 1979).

Given the assumptions of analysis of variance, this equation represents a random effects model for the $\underline{pxi}$ design (Brennan, 1979). Similarly, the linear model for the proportion of items answered correctly on a test is: $X_{\underline{pI}} = \mu+\mu_{\underline{p}}{}_{\sim}+\mu_{\underline{I}}{}_{\sim}+\mu_{\underline{pI}}{}_{\sim}$ where the subscript "$\underline{I}$" equals the average score for a particular sample of items, and all terms are expressed as averages (Brennan, 1979). From the sample sizes and the mean squares generated in the analysis of variance of the $\underline{pxi}$ design, the variance components associated with each of the score effects in the latter equation can be derived. The total of these variances equals the observed score variance. The variance component associated with the effect of person $\underline{p}$, $\sigma_{\underline{p}}{}^2$, is called the universe score variance and is comparable to the true score variance of classical test theory (Brennan, 1979). Similarly, $\sigma^2_{\underline{pI}}$

equals the error variance in classical test theory for a test of length n_i (Brennan, 1979). The variance of mean test scores over all tests, $\sigma^2_{\underline{I}}$, has no comparable statistic in classical test theory (Brennan, 1979). This fact is not surprising since the classically parallel test assumption requires equal test means. Therefore, $\sigma^2_{\underline{I}} = 0$ for classical test theory (Brennan, 1979). Brennan and Kane (1977a) used these variance components to derive indices of dependability for both norm- and criterion-referenced measurement. Although the focus has been on the latter type of measurement, the index of dependability for norm-referenced measurement is also presented below for comparison purposes.

Cronbach et al. (1972) defined the index of dependability for norm-referenced measurement as the ratio of universe score variance to expected observed score variance. This ratio was found to equal:

$$\varepsilon\rho^2 = \frac{\sigma^2_{\underline{p}}}{\sigma^2_{\underline{p}} + \sigma^2_{\underline{pI}}}$$

This index is called the generalizability coefficient and its estimate equals coefficient alpha (Brennan, 1979).

To obtain a mastery testing coefficient, Brennan and Kane (1977a) assumed the major interest is in estimating the difference between a person's universe score and the cut-off, i.e., $\mu_{\underline{p}} - \lambda$ where both terms are expressed as proportions. To estimate this difference, the person's average test score is subtracted from the cut-off, resulting in an error of estimation equal to: $\Delta_{\underline{pI}} = (X_{\underline{pI}} - \lambda) - (\mu_{\underline{p}} - \lambda)$ where $X_{\underline{pI}}$ is the mean observed score of person $\underline{p}$ on test $\underline{I}$, and the other terms are defined as previously (Brennan and Kane, 1977a). Brennan and Kane (1977a) proved that the variance of these

errors, σ^2_{Δ}, equals $\sigma^2_{\underline{I}} + \sigma^2_{\underline{pI}}$ and, then, defined the index of dependability for mastery measurement in terms of expected squared deviations from λ:

$$\Phi(\lambda) = \frac{\varepsilon_{\underline{p}}(\mu_{\underline{p}}-\lambda)^2}{\varepsilon_{\underline{I}}\varepsilon_{\underline{p}}(\underline{X}_{\underline{pI}}-\lambda)^2} = \frac{\sigma^2_{\underline{p}} + (\mu-\lambda)^2}{\sigma^2_{\underline{p}} + (\mu-\lambda)^2 + \sigma^2_{\underline{I}} + \sigma^2_{\underline{pI}}}$$

There are two very important distinctions between $\Phi(\lambda)$ and the generalizability coefficient. First, the true deviation in the former case equals $\sigma^2_{\underline{p}} + (\mu-\lambda)^2$ while it equals $\sigma^2_{\underline{p}}$ for the generalizability coefficient. The first quantity is the same as the true deviation or numerator in Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, while the numerator of the generalizability coefficient is comparable to the true score variance of classical reliability coefficients. Clearly, these similarities and differences are a function of the intended score interpretations, i.e., mastery measurement $(\mu_{\underline{p}}-\lambda)$ versus norm-referenced measurement $(u_{\underline{p}}-\mu)$. Second, the error variance in $\varepsilon\rho^2$ equals $\sigma^2_{\underline{pI}}$, whereas it equals $\sigma^2_{\underline{pI}} + \sigma^2_{\underline{I}}$ in $\Phi(\lambda)$. Obviously, the error variance for $\Phi(\lambda)$ is greater than its counterpart in $\varepsilon\rho^2$ unless all the test means are equal. The generalizability coefficient does not incorporate $\sigma^2_{\underline{I}}$ into the error variance because the test effect adds a constant to every examinee's score resulting in no change in their relative ordering, i.e., no change in the examinees' norm-referenced scores (Brennan, 1979). Since mastery measurement concerns the absolute magnitude of the distance between an examinee's universe score and the cut-off, any effect which increases or decreases this distance for a particular examinee results in an error of measurement (Brennan, 1979). Despite this fact, Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and

Lovett's ANC

because of t

Livingston's

in a domain h

Both ec

formations of

signal/noise

procedure (Br

of the desire

cedure's purp

or the effect

nation (Brenn

noise determi

(Brennan & Ka

this ratio wa

For mastery m

of the signal

over persons

$(\underline{N}(\underline{d}))$ is def

population of

as $\varepsilon_{\underline{p}}\varepsilon_{\underline{I}}(X_{\underline{p}\underline{I}}-$

for mastery t

$$\Phi(\underline{d}) =$$

Lovett's ANOVA based index do not include $\sigma^2_{\underline{I}}$ in their error variance because of their underlying assumption of classic parallelism. Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ are equal when all the possible tests in a domain have equal means (Brennan & Kane, 1977a).

Both $\varepsilon\rho^2$ and $\Phi(\lambda)$ can alternatively be viewed as monotonic transformations of signal/noise ratios (Brennan & Kane, 1977b). The signal/noise ratio indexes the relative precision of a measurement procedure (Brennan & Kane, 1977b). The signal indicates the magnitude of the desired discrimination needed to achieve the measurement procedure's purpose, and the noise represents the magnitude of the errors or the effect of extraneous factors in blurring the desired discrimination (Brennan & Kane, 1977b). The relative sizes of the signal and noise determine whether the desired discrimination can be made (Brennan & Kane, 1977b). Brennan and Kane's derivation (1977b) of this ratio was based upon the principles of generalizability theory. For mastery measurement, the signal is defined as $\mu_{\underline{p}}-\lambda$ and the power of the signal ($\underline{S}(\underline{d})$) equals the expected value of the squared signal over persons or $\varepsilon_{\underline{p}}(\mu_{\underline{p}}-\lambda)^2$. The noise equals $\underline{X}_{\underline{p}\underline{I}}-\mu_{\underline{p}}$. Noise power ($\underline{N}(\underline{d})$) is defined as the expected value of the squared noise over the population of people and samples of items and is expressed as $\varepsilon_{\underline{p}}\varepsilon_{\underline{I}}(\underline{X}_{\underline{p}\underline{I}}-\mu_{\underline{p}})^2$. Combining this information, the signal/noise ratio for mastery tests, $\Psi(\underline{d})$, becomes:

$$\Psi(\underline{d}) = \frac{\varepsilon_{\underline{p}}(\mu_{\underline{p}}-\lambda)^2}{\varepsilon_{\underline{p}}\varepsilon_{\underline{I}}(\underline{X}_{\underline{p}\underline{I}}-\mu_{\underline{p}})^2} \text{ which equals } \frac{\sigma^2_{\underline{p}} + (\mu-\lambda)^2}{\sigma^2_{\underline{I}} + \sigma^2_{\underline{p}\underline{I}}} \text{ (Brennan \& Kane, 1977b).}$$

The index of dependability can now be expressed as:

$$\Phi(\lambda) = \frac{\Psi(\underline{d})}{1 + \Psi(\underline{d})} = \frac{\underline{S}(\underline{d})}{\underline{S}(\underline{d}) + \underline{N}(\underline{d})}$$ (Brennan & Kane, 1977b).

Although intuitively appealing, using signal/noise ratios to express measurement precision has one major drawback from the author's perspective; its upper limit is not one (Brennan, 1979).

Brennan and Kane's Φ. Kane and Brennan (1977) have shown that the quantities "$(\mu - \underline{C}_{\underline{x}})^2$" in $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and "$(\mu-\lambda)^2$" in $\Phi(\lambda)$ equal the expected consistency due to chance factors. The expected chance agreement depends only upon the marginal distribution of scores, not the reliability of the examinee's performance (Kane & Brennan, 1977). Consequently, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ can be large even when scores from the distribution are randomly assigned to examinees on each test administration (Kane & Brennan, 1977).

By subtracting the quantity $(\mu-\lambda)^2$ from both the numerator and denominator of $\Phi(\lambda)$, Kane and Brennan (1977) introduced a coefficient which does take chance agreement into account. This coefficient equals:

$$\Phi = \frac{\sigma^2_{\underline{p}}}{\sigma^2_{\underline{p}} + \sigma^2_{\underline{I}} + \sigma^2_{\underline{pI}}}$$

where each term is defined as in $\Phi(\lambda)$. Note that Φ is the lower limit of $\Phi(\lambda)$ occurring when $\lambda = \mu$ (Brennan & Kane, 1977b).

Φ is a general or multipurpose index of dependability for criterion-referenced measurement (Brennan & Kane, 1977b). Specifically, Φ can serve as a measure of the reliability of examinees' absolute universe scores and/or as a general index of the reliability

of a measurement procedure designed for several decisions using different cut-off scores (Kane & Brennan, 1977).

The signal power ($\sigma^2_{\underline{p}}$) in Φ is the same as that in the generalizability coefficient and is also comparable to the true score variance in classical test theory. Brennan and Kane (1977b) stated that $\sigma^2_{\underline{p}}$ is the appropriate measure of signal power for Φ since it is independent of the cut-off score and is often used as the signal power in physical measurement. The difference between Φ and the generalizability coefficient lies in the definition of the noise power. Within the context of generalizability theory and the general linear model, Brennan (1979) showed that the error variance in using the observed score as a universe score estimate is $\sigma^2_{\underline{I}} + \sigma^2_{\underline{pI}}$. As previously noted, the error term in the generalizability coefficient is $\sigma^2_{\underline{pI}}$. Therefore, Φ is always less than or equal to the generalizability coefficient. "Intuitively, this is a reasonable characteristic of Φ since domain-referenced interpretations of 'absolute' scores are more 'stringent' than norm-referenced interpretations of 'relative scores'" (Brennan, 1979, p. 23).

Related Coefficients. Two additional coefficients employing squared error loss have been proposed within the context of reliability. Harris' index of efficiency ($\mu_{\underline{c}}{}^2$) equals the squared correlation between mastery state, dummy coded as 0 or 1, and the total test score (Harris, 1972b). In analysis of variance terms, $\mu^2_{\underline{c}}$ equals $\underline{SS}_{\underline{b}} / (\underline{SS}_{\underline{b}}+\underline{SS}_{\underline{w}})$ where $\underline{SS}_{\underline{b}}$ and $\underline{SS}_{\underline{w}}$ are the between and within group sum of squares, respectively (Harris, 1972b). Harris (1972b) stated that his coefficient can be interpreted as the ratio of true

score variance to obtained score variance if the group mean is defined as the true score for every subject within the group. Clearly, the validity of such a denotation is questionable since mastery and non-mastery are not typically defined by just one score value. Another problem is that $\mu_{\underline{c}}^2$ is actually a squared point-biserial correlation coefficient and, therefore, uses an inappropriate loss function (squared-error with respect to the mean) and requires the presence of variability.

Similar to Harris' formula, Marshall's index of separation also assesses the extent to which an instrument achieves separation between two groups of people (Marshall, 1976). Assuming that the expected test scores of the knowledgeable group and the not knowledgeable group are the number of test items ($\underline{n}$) and 0, respectively, Marshall (1976) developed the following index:

$$\underline{S}_{\underline{c}} = \frac{1}{\underline{N}}\left[\sum_{\underline{X}\leq\underline{C}_{\underline{x}}} \underline{f}_{\underline{x}}\left(\frac{\underline{C}_{\underline{x}}-\underline{X}}{\underline{C}_{\underline{x}}}\right)^2 + \sum_{\underline{X}>\underline{C}_{\underline{x}}} \underline{f}_{\underline{x}}\left(\frac{\underline{X}-\underline{C}_{\underline{x}}}{\underline{n}-\underline{C}_{\underline{x}}}\right)^2\right]$$

where $\underline{f}_{\underline{x}}$ is the frequency of score $\underline{X}$, $\underline{N}$ equals the number of examinees, and the other terms are as defined previously. Marshall's definition of error is analogous to Harris', i.e., within group variation, and, consequently, using $\underline{S}_{\underline{c}}$ as a reliability measure for mastery testing is not appropriate.

Characteristics of Squared-Error Loss Indices. Given a desire to interpret scores relative to a particular cut-off, Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{x}})$ and Brennan and Kane's $\Phi(\lambda)$ are currently the best squared error loss reliability coefficients. If a coefficient accounting for

chance is desired, Brennan and Kane's Φ is appropriate. The evaluation and interpretation of these indices require an analysis of the factors influencing them. First, as can be seen from the formulas, these three coefficients increase as the norm-referenced reliability increases (Livingston, 1972b). Intuitively, a more accurate true score estimate also implies less error in estimating the distance between the true score and the cut-off. Second, given the previous statements, it is no surprise that lengthening a test increases the value of $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, $\Phi(\lambda)$, and Φ. Livingston (1972b) algebraically proved that the Spearman-Brown prophecy formula applies to $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, and Marshall (1976) empirically supported this derivation. The magnitudes of $\Phi(\lambda)$ and Φ increase since the error terms, $\sigma^2_{\underline{I}}$ and $\sigma^2_{\underline{pI}}$, decrease (Brennan 1979).

Third, although a lack of variability severely affects classical reliability coefficients, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ do not suffer from this limitation (Kane & Brennan, 1977; Livingston, 1972b). When $\sigma^2_{\underline{x}}=0$, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ reduces to $(\mu_{\underline{x}}-\underline{C}_{\underline{x}})^2 / (\mu_{\underline{x}}-\underline{C}_{\underline{x}})^2$ which equals one when $\underline{C}_{\underline{x}}\neq\mu_{\underline{x}}$ and is undefined when $\underline{C}_{\underline{x}}=\mu_{\underline{x}}$ (Livingston, 1972b). The value assumed by $\Phi(\lambda)$ when the variance equals zero is not as clear-cut. However, the important point is $\Phi(\lambda)$ can still equal one given this situation. This lack of dependence upon variability is intuitively reasonable when placed within a signal/noise ratio context. Even if $\sigma^2_{\underline{p}}=0$, the signal will still be easy to detect as long as the distance between the mean and the cut-off is large relative to the noise (Brennan and Kane, 1977b). These statements do not mean the variance has no effect on these coefficients. A change in the variance can induce a change in $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ when the value of the

norm-referenced reliability coefficient is affected (Livingston, 1972b). On the other hand, Φ does depend upon the existence of score variance and equals zero when everyone scores the same (Brennan, 1979).

Fourth, when the cut-off score equals the sample mean, $\hat{K}^2(X,T_X) = r_{11}$ (norm-referenced reliability coefficient) and, given dichotomously scored items, $\hat{\Phi}(\lambda)=\hat{\alpha}_{21}$ (Brennan, 1977; Livingston, 1972b). In this case, all the coefficients use squared error loss with respect to the mean as the loss function. As the cut-off score moves farther away from the mean in either direction, $\hat{K}^2(X,T_X)$ and $\hat{\Phi}(\lambda)$ increase (Brennan & Kane, 1977a; Livingston, 1972b). In other words, the relationship between these coefficients and the cut-off is characterized by a U function with the lowest point occurring when C_X and λ equal the mean. Obviously, $\hat{K}^2(X,T_X) \geq r_{11}$ and $\hat{\Phi}(\lambda) \geq \hat{\alpha}_{21}$ (Brennan, 1979). Correspondingly, Schmitt and Schmitt (1977) found that the average KR-20 over 147 criterion-referenced tests was equal to .53 while the average $\hat{K}^2(X,T_X)$ was .67, and the difference between these coefficients increased as the distance between the mean and the cut-off increased. Likewise, Downing and Mehrens (1978) found that the mean value of $K^2(X,T_X)$ taken over 33 achievement tests was greater than the mean values of KR-20 and KR-21. Livingston (1972b) presented two reasons why the value of $K^2(X,T_X)$ and $\Phi(\lambda)$ should change as the distance between the mean and the cut-off changes: (1) if an individual's obtained score is farther away from the cut-off, his true and obtained scores are more likely to lie on the same side of the cut-off. "Then, if two groups of scores have equal variance and equal reliability in the norm-referenced sense, the group of scores

whose mean is farther away from the criterion score must have the greater criterion-referenced reliability" (p. 18); and (2) a change in the cut-off leads to a different interpretation of scores. Similarly, Brennan and Kane (1977b) viewed an increase in the distance between the mean and the cut-off as an increase in the ability to detect the signal. Others have stated that these coefficients' sensitivity to the relative position of the cut-off is either inappropriate or undesirable (Harris, 1972a, 1973; Shavelson, Block, & Ravitch, 1972).

Shavelson, et al. (1972) believed that the cut-off score's effect on the size of $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ means the latter does not directly reflect the measurement's repeatability. (This argument could also pertain to $\Phi(\lambda)$.) However, given the desire to interpret scores in relation to $\underline{C}_{\underline{X}}$, the difference between the mean and $\underline{C}_{\underline{X}}$ reflects true variance and, therefore, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ does reflect the measurement's consistency (Livingston, 1972c).

Harris (1972a) proved that $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ equals the norm-referenced reliability coefficient computed on pooled data from two populations having equal $\sigma^2_{\underline{T}}$, equal $\sigma^2_{\underline{e}}$, and means equidistant above and below the cut-off. According to Harris (1972a), $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ is deficient because ceiling and floor effects do not always allow one to postulate the existence of two means equidistant from $\underline{C}_{\underline{X}}$. Therefore, the higher reliabilities obtained with $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ are simply due to implicitly increasing the range of talent (Harris, 1972a). In rebuttal, Livingston (1972a) stated, "Criterion-referenced test score interpretations do not require that the criterion score be conceptualized as the mean of some distribution" (p. 9). Simply stated, one must reject

the notion that the first moment of a distribution has to be the mean (Lovett, 1977).

Livingston's coefficient was also criticized because the standard error of measurement remains constant even though $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ increases as the cut-off score moves further from the mean, i.e., the use of the higher $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ as opposed to a classical coefficient does not lead to a more dependable estimate of where a particular examinee truly falls relative to $\underline{C}_{\underline{X}}$ (Harris, 1972a; Shavelson, et al. 1972). This criticism also applies to $\Phi(\lambda)$ (Brennan & Kane, 1977a). However, reliability refers to the dependability of a group of scores, not a single score (Livingston, 1972a). When a mastery decision must be made for every group member, the larger value of $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ implies a more reliable overall estimate of each member's mastery state (Livingston, 1972a). The situation is analogous to the effect that an increase in variance has upon a classical coefficient. Moreover, the standard error of measurement and the squared error criterion-referenced reliability coefficients provide different information: the former measures the variability of an individual's scores independent of the cut-off score, while the latter indicates the consistency of scores relative to the cut-off (Berk, 1980).

In another critique, Harris (1973) showed that the squared standard errors of estimate associated with a linear prediction of true score and of the observed score on a parallel test increase when $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ is substituted for $\underline{r}_{11}$ in the regression equations. Livingston (1973) considered this substitution inappropriate because $\underline{r}_{11}$, not $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$, is the least squares linear regression coefficient. Replacing $\underline{r}_{11}$ by $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ removes the regressor for the

mean and clearly results in an increased residual variance (Livingston, 1973).

Finally, the appropriateness of $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ was questioned because these coefficients increase as the cut-off moves from the mean toward either mode of a symmetric bimodal distribution (Marshall, 1976; Marshall & Serlin, 1979; Subkoviak, 1976). Intuitively, one would expect the opposite to be true, i.e., $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ should be greatest when the mean equals the cut-off since the mean is the point of lowest score concentration and, therefore, the point at which more people should be reliably assigned to mastery states (Marshall, 1976; Subkoviak, 1976). Clearly, this counterintuitive relationship also applies to a unimodal skewed distribution (Marshall & Serlin, 1979). In short, the magnitude of $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ is sensitive to the distance between the mean and the cut-off, but not to the cut-off's relative position to the mode or to heavy score density areas (Marshall & Serlin, 1979). This criticism is unwarranted given that $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ define reliability as the average squared deviation from the cut-off. Like any mean, this average is heavily affected by outliers present in skewed distributions. As the cut-off approaches the mode of such distributions, the outliers become even more influential resulting in an increased average squared deviation. An analogous process occurs for bimodal distributions. In summary, when the cut-off approaches heavy score density areas, more individuals are likely to be misclassified but the reliability in terms of the average squared deviation increases and is appropriately reflected by the magnitude of $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$.

Reliability Formulations Based Upon Threshold Loss

Carver. Carver (1970) introduced two methods for assessing the reliability of mastery measurement. One method consisted of administering the same test to two comparable groups and comparing the percentages of examinees achieving mastery in each group. In the other procedure, the percentages of examinees achieving mastery on two parallel tests are compared. Both procedures are subject to the same limitation; the two percentages compared can be equal even if the measure unreliably classifies every individual (Subkoviak, 1978b). For example, according to the second procedure, perfect reliability can be obtained when 40% of the examinees are classified as masters based on the first test administration and a different 40% are classified as masters on the second administration (Subkoviak, 1978b). Another problem with these procedures is that they do not allow consistent non-mastery decisions to contribute to the reliability measure.

Hambleton and Novick. Hambleton and Novick (1973) suggested that reliability be expressed as the proportion of times a consistent mastery decision is made with two parallel measurement procedures. (Hambleton and Novick (1973) do not use the proportion correct score as an examinee's true score estimate nor as the whole basis for mastery classification. First, they recommend using a Bayesian estimation procedure to determine the probabilities of an examinee being a master and a non-master. Then, based upon the criterion of minimizing threshold loss, these probabilities and the estimated

losses caused by making erroneous decisions are used to classify an examinee.) Given m mastery states, their index can be expressed as:

$$p_o = \sum_{i=1}^{m} p_{ii}$$

where p_{ii} is the proportion of people classified in the ith mastery state on both test administrations (Hambleton & Eignor, 1979). This coefficient is frequently called the coefficient of agreement. Although they certainly were not referring to mastery testing at the time of their writing, Goodman and Kruskal (1954) had suggested using this index as a reliability measure for two polytomies consisting of the same classes.

The upper limit of p_o is, of course, one. Its size is partly a function of the magnitude of the cut-off score relative to the examinees' ability level. For example, p_o will be high when the cut-off score is very low and the examinees have just completed a training program relevant to the tested skill (Millman, 1974). In other words, p_o does not take into account the proportion of agreement expected merely by chance (Kane & Brennan, 1977; Swaminathan et al., 1974). This fact has led to criticism of this index since, as long as the base rate for one category is high, p_o can be high even if the measurement procedure does not contribute to correct classification.

Goodman and Kruskal; Koslowsky and Bailit. In 1954, Goodman and Kruskal advanced an alternative to p_o. They recommended using their index when no relevant continuum underlying the classification scheme existed and when the classifications did not have ordinal properties (Goodman & Kruskal, 1954). One can easily argue that mastery

measurement satisfies neither of these conditions. However, using their measure is possible when only two classifications exist since the ordinal properties are largely irrelevant and since interest lies in evaluating mastery, not an examinee's score on the underlying continuum. The proposed reliability measure is:

$$\lambda_{\underline{r}} = \frac{\sum_{\underline{i}} p_{\underline{i}\underline{i}} - \left[1/2\ (P_{\underline{M}.} + P_{.\underline{M}})\right]}{1 - \left[1/2\ (P_{\underline{M}.} + P_{.\underline{M}})\right]}$$

where $p_{\underline{i}\underline{i}}$ is defined as previously, and $P_{\underline{M}.}$ and $P_{.\underline{M}}$ represent the marginal proportions corresponding to the modal category for rows and columns, respectively. The numerator equals the decrease in the probability of misclassification occurring when an examinee's mastery status is known on one test as opposed to when no information is available (Goodman & Kruskal, 1954). In the latter case, the best guess of the examinee's status is the modal class (Goodman & Kruskal, 1954). The denominator equals the probability of misclassification given no information, and the coefficient equals the proportionate decrease in the probability of misclassification as one moves from the no information situation to a situation where the individual's status is known on one test administration (Goodman & Kruskal, 1954).

Koslowsky and Bailit (1975) expanded upon this formula to determine the reliability of a series of items. This extended index can be used to assess the reliability of a series of mastery decisions. Their measure simply equals the average of $\lambda_{\underline{r}}$ taken over all the mastery decisions ($\underline{a}$):

$$\lambda'_{\underline{r}} = \frac{1}{\underline{N}} \sum_{\underline{a}} \left\{ \frac{\sum_{\underline{i}} p_{\underline{i}\underline{i}} - \left[1/2\ (P_{\underline{M}.} + P_{.\underline{M}})\right]}{1 - \left[1/2\ (P_{\underline{M}.} + P_{.\underline{M}})\right]} \right\}$$

A problem with λ_r and λ'_r is that they are indeterminate when all examinees are masters (non-masters) and both test administrations classify them as such. Clearly, the measurement is perfectly reliable in this case. Koslowsky and Bailit (1975) suggested automatically assigning a value of 1 to λ_r when this situation occurs. Cohen (1960) questioned the appropriateness of λ_r as a reliability index since using the modal category as the "best guess" in the no information situation is more logical within the context of prediction rather than reliability.

Swaminathan, Hambleton, and Algina. To eliminate the influence of chance agreement found with p_o, Swaminathan et al. (1974) proposed using Cohen's coefficient kappa, K. This coefficient is defined as:

$$K = \frac{(p_o - p_c)}{(1 - p_c)}$$

where p_c is the proportion of agreement expected by chance alone or $\sum_{i=1}^{m} p_{i.} p_{.i}$ (Cohen, 1960). The symbols $p_{i.}$ and $p_{.i}$ represent the marginal proportions in a joint classification of the same decision categories on two test administrations, or the proportion of examinees assigned to a mastery state, i, on the first and second test administrations, respectively (Swaminathan et al., 1974). Therefore, p_c is actually a function of the group composition and is the proportion of agreement one would obtain regardless of whether or not the two administrations were statistically independent (Hambleton & Eignor, 1979).

The numerator of K equals the difference between the obtained and the chance proportions of agreement while the denominator equals the maximum value this difference can assume (Millman, 1974). Therefore, K measures the proportion of agreement obtained over and above that expected by chance alone and is, in a sense, independent of the proportion of masters and non-masters in a particular group (Hambleton & Eignor, 1979; Swaminathan et al., 1974).

A limitation of K, as well as of p_o, is that their computation requires two test administrations. Since obtaining data on a parallel test or a retest is not always feasible, an index of classification consistency estimated from a single test administration is definitely needed.

Subkoviak. Subkoviak (1976) offered a single test administration estimate of p_o. He first defined the coefficient of agreement for person "i" as the probability of i being placed in the same mastery state on two parallel tests:

$$p_o^{(i)} = P(X_i \geq C_x, X'_i \geq C_x) + P(X_i \leq C_x, X'_i \leq C_x)$$

where X and X' represent the two test administrations. The first term on the right of the equation denotes the joint probability of person i being consistently classified as a master, and the second term represents the joint probability of a consistent non-mastery decision. Subkoviak then defined the coefficient of agreement (p_o) for a group of N examinees as the mean of the individual $p_o^{(i)}$:

$$p_o = \sum_{i=1}^{N} p_o^{(i)} / N$$

To obtain estimates of $p_o^{(i)}$ from a single test administration, Subkoviak assumed: (1) scores on the two tests were independent for a fixed examinee; and (2) given an individual's true score, the conditional obtained score distributions on both tests were identically binomial. These assumptions led to the following equation for $p_o^{(i)}$:

$$p_o^{(i)} = (P(X_i \geq C_x))^2 + (1-P(X_i \geq C_x))^2 \quad \text{where}$$

$$P(X_i \geq C_x) = \sum_{X_i = C_x}^{n} \binom{n}{X_i} P_i^{X_i} (1-P_i)^{n-X_i}$$

In the latter equation, P_i denotes an individual's true probability of obtaining a correct item response, n equals the number of test items, and X_i represents individual i's obtained score.

Once $P(X_i \geq C_x)$ has been calculated from the data obtained on one test administration, both $p_o^{(i)}$ and p_o can be easily computed. The key to determining $P(X_i \geq C_x)$ is estimating P_i. One could choose the maximum likelihood estimate which equals X_i/n (Subkoviak, 1976). However, the standard error of this estimate is $\sqrt{P_i(1-P_i)/n}$ which is relatively large when $n \leq 40$ (Subkoviak, 1976). Due to this limitation, Subkoviak (1976) recommended using a regression estimate of P_i when the observed scores approximately follow a negative hypergeometric unimodal distribution. Specifically, he proposed the following equation:

$$\hat{P}_i = \left[\alpha_{21} \ (X_i/n)\right] + \left[(1-\alpha_{21}) \ (\mu_x/n)\right]$$

where α_{21} and μ_x equal KR-21 and the test mean, respectively. (In a later paper, Subkoviak (1978b) used KR-20 instead of KR-21 in this equation.) This regression estimate is particularly useful when n is small because the estimate incorporates collateral information

provided by the group (Subkoviak, 1976). The validity of this approach depends upon the sample estimate of the mean and the reliability (Subkoviak, 1976). When the test score distribution is bimodal, Subkoviak (1976) recommended computing separate regression equations for each group or population. Bayesian estimation procedures have also been developed.

Algina and Noe (1978) compared the bias and standard error of $\hat{p}_o$ based upon the maximum likelihood true score estimate to that of $\hat{p}_o$ using the regression estimate. (They used KR-20 rather than KR-21 as the regressor.) Bias and standard error were defined as the mean square deviation of $\hat{p}_o$ from p_o over replications and the standard deviation of this estimate across replications, respectively. The data were simulated using various values for the number of examinees, true score variance, number of items, and cut-off score. Basically, the bias of $\hat{p}_o$ for both true score estimators was affected by the cut-off score, the true score variance, and the number of items. However, changes in these factors affected the extent and/or direction of the biases associated with these two models differently. The regression estimator resulted in a substantially biased estimate when the cut-off scores were close to the mean true score and KR-20 $\geq$.48. In all other cases, the bias was reasonably small. On the other hand, the maximum likelihood estimator tended to result in a substantially biased $\hat{p}_o$ when the cut-off was close to the mean and KR-20 $\leq$ .32. The standard error of $\hat{p}_o$ for both estimators was small in all conditions and increased slightly as the number of examinees decreased. Algina and Noe concluded that, in most cases, using the KR-20 estimator with the binomial error model produced accurate p_o estimates for tests

conforming to this model. However, they also suggested averaging the maximum likelihood and regression p_o estimates when KR-20 is large.

Since Subkoviak's procedure depends upon both an independence and a binomial assumption, the reasonableness of these assumptions should be discussed. The former assumption means the errors of measurement on parallel tests are independent for examinee i and can be met if the tests contain different items or are administered at different times (Subkoviak, 1976). These are obviously the conditions under which many of the classical reliability estimates are determined. When the independence assumption is violated, Subkoviak's index under- or over-estimates the dual administration p_o depending upon whether the two tests are positively or negatively correlated, respectively (Subkoviak, 1976).

To satisfy the binomial assumption, items must be independent and have the same difficulty level. These conditions may not accurately reflect the real world (Gross & Shulman, 1980; Subkoviak, 1976). According to Brennan (1979), items should not be expected to have the same difficulty level. Violating this assumption results in a conservative estimate of mastery classification consistency (Subkoviak, 1976). More accurate p_o estimates can be obtained by replacing the binomial with a compound binomial model which allows varying item difficulties (Subkoviak, 1976). However, Marshall and Serlin (1979) found that these two models produced almost the same results, except in one case.

The binomial error model seems better suited for describing the conditional test score distribution than does the normal error model typically applied in norm-referenced measurement (Brennan, 1974). The

problem with the latter model concerns its assumptions that the errors of measurement are independent of true score and are distributed normally with a mean of zero and homogeneous variances. In criterion-referenced measurement, individuals commonly receive a score of one (expressed as the proportion of correct answers) since they are being trained to achieve mastery (Brennan, 1974). Adopting the classical assumption that $\varepsilon(\underline{E}/\underline{T}) = 0$ implies that people with a true score of one always obtain this score (Lord and Novick, 1968). Likewise, those with a true score equalling zero must always score zero since the observed score can never be negative (Lord & Novick, 1968). In either case, the variance of the errors of measurement is zero.

> This conclusion shows that under any model with bounded observed score and unbiased errors (not all zero), the conditional distribution of the observed score cannot be independent of true score; equally, the conditional distribution of the error of measurement cannot be independent of true score (Lord & Novick, 1968, p. 509).

Consequently, the normality, homogeneity of variance, and independence assumptions of the normal error model are not appropriate for describing the conditional score distribution for criterion-referenced measurement (Brennan, 1974). The formula for the binomial distribution indicates that this model does not make these assumptions.

In summary, Subkoviak's procedure requires the following steps: (1) compute $\hat{\underline{P}}_{\underline{i}}$ through the appropriate regression equation; (2) compute $\underline{P}(\underline{X}_{\underline{i}} \geq \underline{C}_{\underline{x}})$ assuming a binomial distribution of test scores given $\hat{\underline{P}}_{\underline{i}}$; (3) determine $\underline{p}_{\underline{o}}^{(\underline{i})}$; (4) sum the individual $\underline{p}_{\underline{o}}^{(\underline{i})}$ and divide by $\underline{N}$ to obtain $\underline{p}_{\underline{o}}$. If the univariate and bivariate score distributions are approximately normal, the procedure outlined above need not be used to

estimate p_o (Subkoviak, 1976). Subkoviak (1976) proposed the following equation:

$$p_o' = P(X_i \geq C_x, X_i' \geq C_x) + P(X_i < C_x, X_i' < C_x) = 1 - \left[2(P(z < c_x) - P(z < c_x, z' < c_x))\right]$$

where $c_x = (C_x - .5 - \mu_x)/\sigma_x$, σ_x equals the standard deviation of X, and μ_x equals the mean of X (Subkoviak, 1976). In this equation, $P(z < c_x)$ represents the probability that a standardized normal variable is less than c_x and can be found in univariate normal distribution tables. $P(z < c_x, z' < c_x)$ is the probability that two standardized normal variables with a correlation equal to KR-20 are both less than c_x. This probability is obtained from tables of the bivariate normal distribution.

At a later time, Subkoviak (1978b) introduced a single test administration estimate of coefficient kappa by computing the probability of chance agreement which would occur given his model of the data. This probability equals:

$$P_c = 1 - \left\{2\left[\Sigma(\hat{P}(X_i \geq C_x))/N - (\Sigma(\hat{P}(X_i \geq C_x))/N)^2\right]\right\}$$

This formulation was derived by defining the base rate for mastery classification as the average probability (taken across examinees) of being designated a master.

Marshall and Haertel. Marshall and Haertel (1975) also proposed a single administration estimate of p_o, known as coefficient beta (β). Their coefficient equals "the mean of all possible split-half coefficients of agreement" and is, consequently, analogous to coefficient alpha (Marshall & Haertel, 1975, p. 3).

To derive β, scores on a hypothetical $2n$-item test must first be simulated from examinees' scores on an n-item test. Using the binomial error model, this simulation is accomplished via the following equation:

$$N_W = \sum_{X=0}^{n} N_X \binom{2n}{W} (X/n)^W (1-(X/n))^{2n-W}$$

where N_X denotes the frequency of score X on the n-item test, and N_W equals the frequency of score W on a $2n$-item test. Using these simulated scores, Marshall and Haertel define β for an n-item test as:

$$\beta = \frac{1}{\upsilon} \sum_{o=1}^{\upsilon} p_o$$

where p_o is the proportion of agreement consistency between two split-half tests of n items each, and υ is the number of possible splits which can be obtained from the $2n$-item test. The latter quantity equals $\binom{2n}{n}$. Marshall and Haertel's computational formula for β is:

$$\beta = 1/N\left[\sum_{W=0}^{C_X-1} N_W + \sum_{W=C_X}^{2C_X-2} N_W \cdot \Phi_W(W-(C_X-1),\, C_X-1) + \sum_{W=2C_X}^{n+C_X-1} N_W \cdot \Phi_W(C_X, W-C_X) + \sum_{W=n+C_X}^{2n} N_W\right]$$

where

N = number of examinees

W = examinee's score on a $2n$-item test

n = number of test items

N_W = frequency of score W

C_X = cut-off score on an n-item test

$\Phi_W(a,b) = \sum_{j=a}^{b} \binom{W}{j}\binom{2n-W}{n-j} / \binom{2n}{n}$ or the proportion of splits

resulting in a half-test score of from a to b inclusive, given a total score of W.

As can be seen, β is the mean of its additive parts and, therefore, each examinee's score makes a specific contribution to beta's magnitude (Marshall, 1976). The further the score departs from the value $2C_x-1$, the more it contributes to the size of β (Marshall, 1976). A score equalling $2C_x-1$ makes a zero contribution; at this particular value, the examinee must always be classified as a master on one half of the test and a non-master on the other half (Marshall, 1976).

Similar to Subkoviak's model, the validity of using the binomial error model in Marshall and Haertel's formula is questionable. However, results of a study investigating the bias of various estimates showed that β produced quite accurate estimates of p_o when items were not homogeneous, particularly for longer tests (n=30, n=50) (Subkoviak, 1978a).

One drawback of this model, as noted in a personal communication from Marshall (1980), is the use of the proportion correct score as the true score estimate in computing N_W. As previously mentioned, the standard error of this estimate is reasonably large when $n \leq 40$. Apparently, Marshall no longer recommends this procedure (Marshall & Serlin, 1979). A regression or Bayesian estimate can easily be incorporated into the procedure. In one study, Marshall and Serlin (1979) actually used a predictive Bayesian beta model as well as other models to obtain the frequency distribution for a $2n$-item test.

Huynh. Huynh (1976) developed a single administration estimate of p_o and kappa based upon Keats and Lord's beta-binomial test score model. Like Subkoviak's and Marshall and Haertel's formulations, this model assumes an examinee's scores given his/her true score follow a binomial distribution (Huynh, 1976; Keats & Lord, 1962). According to Huynh (1976), assuming similarity of item difficulty and item content (i.e., item exchangeability) is reasonable for criterion-referenced measurement because all items should measure a single trait. Moreover, his p_o appears robust with respect to violation of the former assumption (Subkoviak, 1978a). Specifically, violation of this assumption resulted in slightly conservative estimates of reliability for a 10-item test and had little effect on longer tests (Subkoviak, 1978a).

The Keats-Lord model also assumes true scores follow a beta distribution. The beta distribution family includes a wide range of shapes although multi-humped distributions are not included (except for a U-shaped function where the modes occur at 0 and n). The parameters of the beta distribution, α and β, can be computed from the mean and standard deviation of a large sample score distribution:

$$\alpha = \left(-1 + \frac{1}{\alpha_{21}}\right) \cdot \mu_x$$

$$\beta = -\alpha - n + \frac{n}{\alpha_{21}} \quad \text{where } \alpha_{21} = KR\text{-}21 = \frac{n}{n-1}\left[1 - \frac{\mu_x(n-\mu_x)}{n\sigma^2}\right] \text{(Huynh, 1976).}$$

Under the beta-binomial model, the observed score distribution has a negative hypergeometric distribution with the following density:

$$f(x) = \frac{\binom{n}{x} B(\alpha+x,\ n+\beta-x)}{B(\alpha,\beta)}$$

where $\underline{B}$ denotes the beta function (Huynh, 1976). Huynh (1976) has provided computational formulas for evaluating $\underline{f}(\underline{x})$. Estimating $\underline{p}_o$ and kappa also requires determining the joint distribution of equivalent test forms, $\underline{f}(\underline{x},\underline{y})$. Assuming local independence with respect to the true score, $\underline{f}(\underline{x},\underline{y})$ can be simulated. This distribution follows a bivariate negative hypergeometric or beta-binomial distribution with the following density:

$$\underline{f}(\underline{x},\underline{y}) = \frac{\binom{\underline{n}}{\underline{x}}\binom{\underline{n}}{\underline{y}}}{\underline{B}(\alpha,\beta)}\underline{B}(\alpha+\underline{x}+\underline{y},\ 2\underline{n}+\beta-\underline{x}-\underline{y}) \quad \text{(Huynh, 1976).}$$

Huynh (1976) also presented computational formulas for $\underline{f}(\underline{x},\underline{y})$.

Given a particular cut-off score, these formulas can be used to calculate the proportion of examinees who would be placed in the mastery category on both test forms ($\underline{p}_{11}$), the proportion who would be consistently classified as non-masters ($\underline{p}_{00}$), and the proportion who would be given mastery status by only one form ($\underline{p}_1$). These proportions are defined in the following manner:

$$\underline{p}_{11} = \sum_{\underline{x},\underline{y}=\underline{C}_{\underline{x}}}^{\underline{n}} \underline{f}(\underline{x},\underline{y})$$

$$\underline{p}_{00} = \sum_{\underline{x},\underline{y}=0}^{\underline{C}_{\underline{x}}-1} \underline{f}(\underline{x},\underline{y}) \text{ and}$$

$$\underline{p}_1 = \sum_{\underline{x}=\underline{C}_{\underline{x}}}^{\underline{n}} \underline{f}(\underline{x}) \qquad \text{(Huynh, 1976).}$$

Given the assumption that the marginal distribution is the same for each form, Huynh (1976) defined $\underline{p}_o$ and kappa as:

$$\underline{p}_o = \underline{p}_{11} + \underline{p}_{00} \text{ and } \underline{K} = \frac{\underline{p}_{11} - \underline{p}_1^2}{\underline{p}_1 - \underline{p}_1^2}$$

When the cut-off score is small, the following formula for $\underline{K}$ is far more convenient:

$$\underline{K} = \frac{\underline{p}_{00} - \underline{p}_0^2}{\underline{p}_0 - \underline{p}_0^2}$$

where $\underline{p}_0$ is the proportion of examinees classified as non-masters by only one test form (Huynh, 1976).

When the number of test items is moderately large (e.g., $\underline{n} > 10$), Huynh (1976) suggested using a normal approximation procedure to estimate kappa. In this procedure, an arcsine transformation is applied to the data, resulting in an approximately normal score distribution. Univariate and bivariate normal distribution tables are then used to estimate the probabilities needed for computing $\underline{K}$.

Peng and Subkoviak (1980) found that, in the vast majority of his simulated distributions, a simple normal approximation procedure using Yate's correction resulted in less proportionate error in estimating $\underline{K}$ than did Huynh's normal approximation procedure. Peng varied the beta distribution parameters, the cut-off score, and the test length. The upper limit of the latter variable was 30. Using real data, Peng (1979) collaborated his findings. The superiority of the simple normal procedure was more pronounced for short tests and/or moderate cut-off scores (between 65% and 85%). Similar results were obtained when the two normal approximation procedures were used to estimate $\underline{p}_0$ (Peng, 1979; Peng & Subkoviak, 1980).

Characteristics of Threshold Loss Indices. As can be seen, the most appropriate threshold loss coefficients are divided into two categories: (1) $\underline{p}_0$ coefficients; and (2) kappa coefficients. Because

the former indices do not take account of chance agreement while the latter ones do, various population and test characteristics affect p_o and kappa differently. Since research has shown these factors affect dual and single administration coefficients similarly, the following discussion applies to both unless otherwise stated.

First, under the assumption of exchangeability, the theoretical lower limit of p_o is the proportion of agreement expected by chance, while kappa's limit is zero (Huynh, 1978; Subkoviak, 1978b). In general, however, the lower limit of kappa, computed from two test administrations, depends upon the marginal distributions (Cohen, 1960). The upper limit of both coefficients is +1.00.

Second, as the cut-off approaches the extremes, p_o generally approaches one (Marshall, 1976; Marshall & Haertel, 1975; Subkoviak, 1976, 1977). This trend is particularly evident for symmetric unimodal distributions (Marshall & Haertel, 1975). On the other hand, kappa generally approaches its lowest value as the cut-off moves toward the distribution extremes (Huynh, 1976; Subkoviak, 1977). This difference can be partly explained by the fact that the probability of chance consistency generally tends toward one as the cut-off approaches the extremes (Huynh, 1976). Therefore, p_o also approaches one, while kappa decreases because not much opportunity exists for increasing agreement above chance (Huynh, 1976).

Third, the magnitude of p_o has been found to increase as the distance between the cut-off and areas of heavy score density (e.g., the mode) increase (Eignor & Hambleton, 1979; Marshall, 1976; Subkoviak, 1976, 1977). Given $r_{tt} < 1.00$, examinees scoring close to

the cut-off on the first test administration could easily obtain a score on the opposite side of the cut-off on the second administration. On the other hand, those further away from the cut-off would more likely be placed in the same mastery state in both testing sessions. Therefore, the greater the number of scores further away from the cut-off, the higher the $\underline{p}_{\underline{o}}$. Exceptions to this relationship have been found for the single administration coefficients (Marshall & Serlin, 1979). Marshall and Serlin (1979) examined the behavior of these coefficients given five different distributions: (1) bell-shaped; (2) highly negatively skewed unimodal; (3) bimodal with a stronger mode at the higher end; (4) symmetric bimodal with modes widely separated; and (5) symmetric bimodal with modes close together. With the exception of the fifth distribution, the size of Subkoviak's $\hat{\underline{p}}_{\underline{o}}$ generally reflected the distance between the cut-off and the mode for both unimodal and bimodal distributions. Fortunately, the fifth distribution is atypical in mastery testing (Marshall & Serlin, 1979). Huynh's $\hat{\underline{p}}_{\underline{o}}$ reflected the cut-off's position for unimodal distributions and bimodal distributions with extreme modes, but not for bimodal distributions not belonging to the beta-binomial family. For Marshall and Haertel's index, five different test score models were used to simulate scores on a $2\underline{n}$-item test from scores on an $\underline{n}$-item test. The adequacy of their $\hat{\underline{p}}_{\underline{o}}$ in reflecting the cut-off's relative position depended upon the model used to generate scores. One of the best models was a binomial regression model comparable to that used in Subkoviak's index. This model produced results similar to those obtained with Subkoviak's $\hat{\underline{p}}_{\underline{o}}$. An averaged double binomial model introduced by Marshall and

Serlin also reflected the location of the mode(s) for both unimodal and bimodal distributions.

In contrast, given the assumption of exchangeability, Huynh (1978) mathematically proved that kappa is an inverted U function of the cut-off when the data are normally distributed. This relationship was also empirically supported for normally distributed data as well as for various beta-binomial and some bimodal distributions (Eignor & Hambleton, 1979; Huynh, 1976, 1978; Marshall & Serlin, 1979; Subkoviak, 1977). Apparently, the location of the cut-off relative to the score density affects kappa in a manner opposite to its effect on $\underline{p}_o$, i.e., kappa is greater when the cut-off is located near heavy score density areas. Intuitively, one might expect kappa to behave similarly to $\underline{p}_o$. The difference appears to be due once again to the influence of chance agreement. Specifically, in many distributions, $\underline{p}_c$ decreases as the cut-off approaches heavy score density areas, leading to a decrease in $\underline{p}_o$. However, kappa increases because more opportunity exists for agreement above that expected by chance.

Generally, the cut-off score appears to affect the magnitude of $\underline{p}_o$ and kappa in two ways, i.e., through its relative position to the extremes and to the heavy score density areas. Conceivably, these two influences could interact, producing some unpredictable results. For example, what would happen to the size of $\underline{p}_o$ and kappa if the cut-off and the mode were equal to $\underline{n}$? Marshall (1976) used this interactional effect to explain the unpredictable relationships found between the cut-off and his coefficient. This effect probably also explains some unforseen trends Eignor & Hambleton (1979) found with kappa.

Fourth, $\underline{p}_o$ does not require score variability to attain its upper limit but kappa does (Kane & Brennan, 1977). However, both coefficients increase as the variance increases (Huynh, 1976; Marshall, 1976; Swaminathan et al., 1974). A large variance implies extreme scores and, consequently, better differentiation between masters and non-masters (Marshall, 1976).

Fifth, although all the aforementioned variables affect $\underline{p}_o$ and kappa differently, the test length and the classical reliability coefficients affect them similarly. Specifically, as the number of test items increase, $\underline{p}_o$ and kappa increase (Eignor & Hambleton, 1979; Huynh, 1976, 1978; Marshall, 1976; Marshall & Haertel, 1975; Subkoviak, 1978b; Swaminathan et al., 1974). Increasing the test length probably results in a more accurate true score estimate and, consequently, a more reliable estimate of an examinee's mastery state. Correspondingly, as the classical reliability coefficient increases so should $\underline{p}_o$ and kappa. Marshall (1976) found the mean of his coefficient taken over various cut-off scores was highly correlated with KR-21 across several distributions (Rho=.93). Given parallel tests, dual administration kappa was mathematically and empirically shown to increase as the classical reliability coefficient increased for a normal distribution and a beta-binomial model, respectively (Huynh, 1978). In addition, Downing and Mehrens (1978) found that Huynh's single administration kappa coefficient correlated .96 and .98 with KR-20 and KR-21, respectively. On the other hand, Algina and Noe's results (1978) did not support a relationship between Subkoviak's $\hat{\underline{p}}_o$ and a classical coefficient.

Synthesis

In the foregoing discussion, which coefficient to use in a particular mastery testing situation was not delineated. The present section addresses this issue by synthesizing the previous material and determining the major distinctions among the various coefficients.

Using the concept of agreement functions, Kane and Brennan (1977) provided a single consistent framework for viewing the reliability coefficients. As explained by Kane and Brennan (1977), an agreement function denotes the extent of agreement between the interpretation of examinees' scores on randomly parallel tests. For mastery measurement, coefficients are based upon either a squared-error (with respect to the cut-off) or a threshold agreement function corresponding to the squared-error and threshold loss functions previously discussed. Kane and Brennan showed that the indices equal either the proportion of maximum agreement achieved by the measurement procedure or the proportion of maximum agreement achieved over and above that expected by chance. Maximum agreement is the expected agreement between a testing procedure and itself, while the agreement produced by the measurement procedure is the expected value of the agreement function. Figure 2 presents the major single test administration reliability coefficients within their appropriate categories, formed by crossing type of agreement function with the presence of a chance agreement correction.

One must first decide whether to use squared error or threshold agreement coefficients (Kane & Brennan, 1977). Since the former coefficients are concerned with the extent of deviation from the cut-off, their size reflects the magnitude of errors (Brennan & Kane,

Type of Agreement Function	Chance Agreement: Uncorrected	Chance Agreement: Corrected
Squared Error	Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ Brennan & Kane's $\Phi(\lambda)$	Brennan & Kane's Φ
Threshold	Subkoviak's p_o Marshall & Haertel's $\underline{p}_o$, Huynh's $\underline{p}_o$	Subkoviak's kappa Huynh's kappa

Figure 2.--Mastery Testing Reliability Formulations

1977a). In other words, they do not consider all inconsistent classifications or misclassifications to be equally serious, but assume that misclassifying an examinee whose true ability level is far from the cut-off is much more serious than misclassifying someone whose true ability is close to the cut-off (Brennan & Kane, 1977a). This advantage is particularly compelling since cut-off scores are, to some extent, arbitrarily determined and, therefore, a sharp distinction between masters and non-masters seldom exists (Brennan & Kane, 1977a; Glass, 1978). Furthermore, different procedures for setting cut-off scores result in different cut-offs (Brennan & Lockwood, 1979). However, a drawback of these coefficients is their sensitivity to all errors, even those not resulting in inconsistent mastery decisions (Brennan & Kane, 1977a).

On the other hand, threshold agreement indices do not reflect the magnitude of errors but are only sensitive to errors resulting in misclassification (Brennan & Kane, 1977a). The disadvantage of these coefficients is that they consider all misclassifications to be equally serious (Brennan & Kane, 1977a).

Clearly, neither the squared error nor the threshold agreement coefficients are optimal in every situation. Kane and Brennan (1977) suggested the following course of action:

> The threshold agreement coefficient is appropriate whenever the only distinction that can be made usefully is a qualitative distinction between masters and non-masters. If, however, different degrees of mastery and non-mastery exist to an appreciable extent, the threshold agreement function is not appropriate because it ignores such differences (p. 40).

Since reliability is relative to the score interpretation, the appropriate agreement function should be dictated by the way the scores will be used (Popham & Husek, 1969; Subkoviak, 1978b). If the degree of mastery or non-mastery is of interest, coefficients incorporating a squared-error agreement function are more suitable (Subkoviak, 1978b). This situation occurs when different actions or programs are to be initiated based on how far from the cut-off an examinee scores and/or when distance from the cut-off leads to unequal misclassification losses (Brennan & Kane, 1977a; Popham & Husek, 1969). When only two courses of action are possible and misclassification losses are considered equal, threshold agreement coefficients should be applied (Brennan & Kane, 1977a). Likewise, if there exist more than two mastery categories and no differential misclassification loss related to distance, threshold agreement indices can be used. However, Kane and Brennan (1977) stated that threshold agreement coefficients are inappropriate when more than two mastery classifications exist and these categories are ordered. Addressing the ordered case, Goodman and Kruskal (1954) proposed two other measures which account for how different an individual's mastery classification on two test administrations is. No single

administration index of these ordered coefficients has been formally developed. However, it seems the single administration threshold agreement indices could easily be adapted to this purpose.

The next decision one must face is whether or not to use a coefficient accounting for chance agreement. Differentiating between corrected and uncorrected coefficients is important because they provide different kinds of information about reliability (Kane & Brennan, 1977; Subkoviak, 1978b). The uncorrected squared-error and threshold agreement indices indicate the reliability of the deviation scores and the mastery classifications, respectively, i.e., the consistency of the score interpretation (Kane & Brennan, 1977). Both chance agreement and the consistency contributed by the testing procedure affect the value of these coefficients (Kane & Brennan, 1977). In comparison, corrected coefficients measure only the latter source of consistency, i.e, the contribution of the testing procedure to the reliability of scores over and above that expected by chance (Kane & Brennan, 1977). Clearly, the choice between corrected and uncorrected coefficients depends upon whether one wants to determine the consistency of scores regardless of the causes of this consistency (i.e., test procedure, group composition, group's mean ability) or the reliability of the testing procedure irrespective of the group's characteristic ability or mastery level (Subkoviak, 1977).

In discussing threshold loss indices, Livingston and Wingersky (1979) and Berk (1980) do not recommend using the corrected coefficient, kappa, in situations where an absolute cut-off has been

established because the correction for chance takes the marginal frequencies as given. As stated by Livingston and Wingersky (1979):

> 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 it happened to be (p. 250).

However, the present author fails to see how this fact differentiates kappa from any other reliability estimate which uses sample statistics (e.g., the sample mean) as estimates of population values.

The corrected indices, coefficient kappa and Φ, could be criticized because they approach or equal zero when little or no true mastery score variability exists (i.e., when everyone is placed in the same mastery state or receives the same domain score, respectively) even though the scores may be perfectly reliable (Berk, 1980). However, this criticism is unwarranted. These coefficients' low values in the presence of small variability do not indicate that the mastery scores are unreliable, but simply that the testing procedure does not add much more reliability to the scores above that achieved by chance processes (Kane & Brennan, 1977). In other words, a testing procedure resulting in some sort of criterion-referenced score interpretation must produce variability in terms of those scores if the procedure is going to contribute to reliability (Kane & Brennan, 1977). On the other hand, the uncorrected coefficients can be large even when no true score variability exists because of the score consistency contributed by chance processes. These observations provide a new perspective on Popham and Husek's disagreement with Woodson over the variability issue (Kane & Brennan, 1977). To

reiterate, Popham and Husek contended that variability is not a necessary characteristic of a good criterion-referenced test, while Woodson argued that a test with no variability provides no information. It appears that Popham and Husek's argument applies to the score interpretation, while Woodson's argument applies to the test's contribution to this interpretation (Kane & Brennan, 1977).

As previously discussed, the four types of coefficients depicted in Figure 2 react differently to the relative position of the cut-off. Obviously, the cut-off's location does not affect the corrected squared error coefficient. However, the uncorrected squared error indices are sensititve to the distance between the mean and the cut-off; they increase as this distance increases. The $\underline{p}_o$ and kappa indices are generally not expected to be sensitive to this difference unless the mean reflects heavy score density areas.

On the other hand, squared error indices are not sensitive to the distance between the cut-off and the mode or heavy score density areas, while uncorrected threshold indices are hypothesized to increase as this distance increases. In contrast, the corrected threshold indices appear to be greater when the cut-off is located in heavy score density areas. For example, when scores are normally distributed, a U function characterizes the relationship between the cut-off score and $\underline{p}_o$, while kappa is an inverted U function of the cut-off.

Similar to $\underline{p}_o$, the uncorrected squared error indices are also a U function of the cut-off score given a normal distribution since the mean equals the mode (Marshall, 1976). However, when the distribution is skewed and/or bimodal, uncorrected squared error coefficients will

increase while uncorrected threshold indices will decrease as the cut-off moves from the mean toward the mode(s). Correspondingly, for bimodal distributions, Marshall (1976) found that the magnitude of his $\hat{\underline{p}}_{\underline{o}}$ and $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ did not fluctuate similarly as the cut-off score varied. This observation is particularly relevant in mastery measurement since the score distribution on any given test administration is often bimodal and, in some cases, is expected to be skewed (e.g., after an instructional program) (Marshall, 1976; Marshall & Serlin, 1979).

Although the list of applicable coefficients can be reduced by choosing an appropriate agreement function and deciding whether or not to correct for chance processes, one must still select among alternative formulas in many cases. The choice of an appropriate index in these instances depends upon the number of feasible test administrations, the satisfaction of the assumptions underlying a particular index, the coefficient's robustness to violations of these assumptions, the coefficient's bias in estimating the dual administration population index, and the degree of sampling fluctuation exhibited by the coefficient. In most situations, two test administrations are not possible and, therefore, the applicable coefficients are typically those requiring only one test administration.

If one has decided to use an uncorrected squared error coefficient, one can choose Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and/or Brennan and Kane's $\Phi(\lambda)$. A major difference between these indices is that $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ is based upon classical test theory, while $\Phi(\lambda)$ is derived from

generalizability theory (Brennan, 1979). The latter theory has two distinct advantages over the former. First, generalizability theory provides the opportunity to examine the reliability of data derived from different types of experimental designs, e.g., nested design (Brennan, 1978). This theory also allows one to take account of whether the various effects are fixed or random (Brennan, 1978). Second, generalizability theory can differentiate norm- from criterion-referenced measurement by distinguishing between different error variances, while classical test theory cannot (Brennan, 1979). Specifically, Brennan and Kane's approach indicates that $\sigma^2_{\underline{pI}}$ is the appropriate error term for norm-referenced measurement, while $\sigma^2_{\underline{pI}} + \sigma^2_{\underline{I}}$ is the proper error variance in criterion-referenced measurement (Brennan, 1979). Clearly, the classically parallel test assumption obviates the existence of $\sigma^2_{\underline{I}}$. Generalizability theory assumes tests are randomly parallel. Brennan (1979) finds the classically parallel test assumption unreasonable for criterion-referenced testing since the test construction method does not require content specialists to include only items with the same difficulty level in the domain. If, as expected, the items in a domain have various difficulty levels, it would be very unlikely for all tests constructed from this domain to be classically parallel (Brennan, 1979). Furthermore, since $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ equals $\Phi(\lambda)$ when test means are equal, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ is really a special case of $\Phi(\lambda)$. For these reasons, the more general $\Phi(\lambda)$ appears preferable to $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ (Brennan, 1979). Unfortunately, no empirical research concerning the bias and sampling fluctuation of these coefficients exists.

When considering uncorrected threshold loss indices, the appropriateness of several alternative p_o formulas must be evaluated. If feasible, p_o can, of course, be estimated from two test administrations. The dual administration $\hat{p}_o$ is unbiased and its standard error equals $(p(1-p)/N)^{1/2}$ (Huynh & Saunders, 1979). Generally, formulas for evaluating the standard error of the single administration p_o estimates have not been developed and very little empirical evidence pertaining to their bias and standard error have been produced. However, assuming a beta-binomial score distribution, Huynh (1978) showed that his $\hat{p}_o$ index is asymptotically unbiased and also presented a formula for the asymptotic standard error of this estimate. In addition, Huynh and Saunders (1979) found that Huynh's $\hat{p}_o$ generally underestimated the dual administration p_o for large data sets not conforming to the beta-binomial model as well as for small and moderate sized samples (N=20,40,60). In the former case, the average amount of bias was -2.3% across various test lengths and cut-off scores. For the small and moderate sized samples, the average degree of bias was -2.6% across various test lengths.

Assuming a large sample size, Huynh and Saunders (1979) also compared the standard error of the dual administration $\hat{p}_o$ to that of Huynh's estimate for various beta-binomial distributions, test lengths, and cut-off scores. The mean and KR-21 of the distributions were chosen to reflect one of the following shapes: (1) U-shaped with the higher mode at the upper end of the distribution; (2) symmetric; (3) unimodal with the mode lying between μ and n; and (4) J-shaped. In every instance, the standard error of Huynh's estimate was lower than its dual administration counterpart. On the average, the

stan

smal

distr

sever

Saund

error

admir

all t

(Subk

upon

propo

Marsh

stude

the p

(10,

and s

the n

error

the d

the e

distar

the fa

popula

spondi

admini

error

standard error of the former was 59.3% of the latter. The uniformly smaller standard error of Huynh's $\hat{p}_o$ was also found for large sample distributions significantly different from the beta-binomial in several instances and for small to moderate sized samples (Huynh & Saunders, 1979). Over all the situations considered, the standard error of Huynh's $\hat{p}_o$ was 50.4% and 51.4% of that of the dual administration estimate, respectively.

Only one study correctly compared the bias and standard error of all the p_o estimates (including the dual administration $\hat{p}_o$) (Subkoviak, 1978a). In this study, Subkoviak's coefficient was based upon a compound binomial instead of a binomial model, and the proportion correct score was used as the true score estimate in Marshall's $\hat{p}_o$. Each estimate was computed for 50 random samples of 30 students each and compared to the dual administration p_o obtained in the population (N=1586). Comparisons were made for three test lengths (10, 30, 50) and four cut-off scores (.5n, .6n, .7n, .8n). The mean and standard deviation of each estimate across the 50 samples provided the necessary data for judging the estimate's bias and standard error. All the estimates became more accurate as the test length and the distance between the mean and the cut-off increased. Moreover, the estimates' standard errors decreased. The influence of the distance between the mean and the cut-off can be partly explained by the fact that estimates become more accurate and less variable as the population parameter becomes more extreme (Subkoviak, 1978). Corresponding to Huynh and Saunder's results (1979), the dual administration estimate was unbiased but had the largest standard error regardless of the test length and cut-off score. Huynh's $\hat{p}_o$

underestimated $\underline{p}_o$ for short tests and was, generally, less variable than the other indices for 30- and 50-item tests. Marshall and Haertel's $\hat{\underline{p}}_o$ was biased upward when the cut-off was near the mean and biased downward when the cut-off was in the tails of the distribution. This effect was more pronounced for shorter tests. Conversely, for short tests, Subkoviak's $\hat{\underline{p}}_o$ underestimated $\underline{p}_o$ when the cut-off was near the mean and overestimated $\underline{p}_o$ for more extreme cut-offs. This finding was similar to that found by Algina and Noe (1978). The opposite reaction of Marshall's & Subkoviak's indices may have been due to the use of different true score estimates. Specifically, Marshall and Haertel's use of the proportion correct score produces an overestimate of the true score variance, while Subkoviak's regression true score estimate results in an underestimate of this variance (Algina & Noe, 1978). It should be noted that Subkoviak's $\hat{\underline{p}}_o$ showed no consistent pattern for longer tests. Finally, Marshall and Haertel's index was the least variable but the most biased for $\underline{n}$=10. Except in this latter case, none of the four coefficients was substantially biased.

In evaluating which single administration $\underline{p}_o$ estimate to apply, the assumptions underlying each of them should be examined. All assume the distribution of an examinee's test scores given his/her true score is binomial. Recognizing that the equal item difficulty assumption might be unrealistic, Subkoviak (1976) proposed using the compound binomial instead of the binomial model. However, whether or not this more complicated procedure improves estimation of $\underline{p}_o$ is highly questionable. Use of the compound binomial in Subkoviak's $\hat{\underline{p}}_o$ generally produced results similar to those obtained using the

binomial model (Marshall & Serlin, 1979). Furthermore, Huynh and Saunders (1979) found the standard deviation of item difficulties was not related to the degree of bias associated with Huynh's $\hat{p}_o$ and Huynh's kappa estimate, and Subkoviak (1978a) provided evidence that all three coefficients are robust with respect to violation of the equal item difficulty assumption.

Another assumption implicit in all three single administration coefficients is classic parallelism (Kane & Brennan, 1977). The validity of this assumption in criterion-referenced testing has already been questioned. When tests are not classically parallel, these coefficients will probably overestimate p_o. To the author's knowledge, no empirical evidence addressing this question exists. Those few studies examining the bias of one or more of these estimates included only parallel tests (for example, Subkoviak, 1978a).

Huynh and Saunders (1979) noted that Subkoviak's procedure and Huynh's procedure assume the score distribution is beta-binomial. Therefore, they should have similar patterns of bias and standard error. Huynh and Saunders (1979) concluded that such was the case in Subkoviak's investigation (1978a). Although not explicitly stated, Subkoviak's study of bias appears to have been performed on data following a normal distribution. The normal distribution is not a member of the beta-binomial family, although this family does include a "normally" shaped distribution (Gross & Shulman, 1980). The bias and standard error of these estimates have not been investigated for distributions more typically found in criterion-referenced measurement, i.e., skewed and bimodal (Marshall, 1976; Marshall &

Serlin, 1979). Examining these coefficients given the latter distribution would be particularly interesting because the beta-binomial family does not include bimodal distributions, except for U-shaped and J-shaped functions (Gross & Shulman, 1980). Both these distributions are not expected to occur in the real world (Marshall & Serlin, 1979). Subkoviak (1976, 1978a) has stated that using a single regression equation to estimate the true score in his procedure is inappropriate given a bimodal distribution and has recommended using Huynh's procedure. However, Marshall and Serlin (1979) found that the magnitude of Huynh's $\hat{p}_o$ did not reflect the location of the modes for bimodal distributions, while Subkoviak's $\hat{p}_o$ reflected the mode(s) for both unimodal and bimodal distributions. Although not explicitly stated, the researchers appear to have used a single regression equation to obtain Subkoviak's true score estimate for the bimodal as well as the unimodal distributions. Gross and Shulman (1980) investigated the robustness of the beta-binomial model; they compared empirical values of p_o obtained from two test administrations to the theoretical values of p_o derived from the beta-binomial model when its underlying assumptions were violated. They found that the theoretical and empirical values were in close agreement. However, the authors did not indicate the shape of the score distribution nor how severely the assumptions were violated.

One of the most enlightening findings concerning the p_o estimates evolved from Marshall and Serlin's study (1979). They used five versions of Marshall and Haertel's $\hat{p}_o$ varying in terms of the model used to simulate scores on a $2n$-item test. They found that Huynh's $\hat{p}_o$ and Subkoviak's $\hat{p}_o$ were empirically equivalent to Marshall and

Haerte
applie
Lord b
for Ma
cases.
simulat
summary
other t
lated t
Therefo
choosin
differe
empiric
least b

Fi

index,
Subkovi
kappa e
However
both sm
present
error.

Gi

single a
and pres
several
from a b

Haertel's estimate when the assumptions of the former indices were applied to the latter coefficient. Specifically, when the Keats and Lord beta-binomial model was used to simulate scores on a $2n$-item test for Marshall's $\hat{p}_o$, this index was equal to Huynh's $\hat{p}_o$ in each of 300 cases. Similarly, when a binomial regression model was used to simulate scores, Marshall's and Subkoviak's indices were equal. In summary, Marshall's $\hat{p}_o$ appears to be a general index subsuming the other two coefficients and is equal to them when the data are postulated to meet certain assumptions (Marshall & Serlin, 1979). Therefore, a choice among the three coefficients seems reduced to choosing among various test models rather than among three entirely different coefficients (Marshall & Serlin, 1979). Clearly, much more empirical research is needed to choose which test model results in the least bias and standard error given a particular type of distribution.

Finally, if the situation demands a corrected threshold agreement index, one can use a dual test administration kappa estimate, Subkoviak's model, and/or Huynh's procedure. The dual administration kappa estimate is asymptotically unbiased (Huynh & Saunders, 1979). However, Huynh and Saunders (1979) found a small negative bias for both small (N=20, 40) and moderate (N=60) sized samples. They also presented a formula for computing this estimate's asymptotic standard error.

Given a beta-binomial distribution, Huynh (1978) showed that his single administration kappa formula is also asymptotically unbiased and presented a formula for its asymptotic standard error. For several large data sets, some of which were significantly different from a beta-binomial distribution, Huynh and Saunders (1979) found

that this estimate tended to underestimate the population dual administration kappa. Across various test lengths and cut-off scores, the average percent of bias was -7.8. The same trend was found for small and moderate sized samples; across various test lengths, the average percent of bias was -11.0.

Huynh and Saunders (1979) also compared the standard error of Huynh's kappa to that of the dual administration estimate. Over various beta-binomial distributions, test lengths, and cut-off scores, the standard error of Huynh's kappa was consistently lower. On the average, it was 53.2% of the standard error of the dual administration kappa. The uniformly smaller standard error of Huynh's estimate was also found for large data sets with distributions significantly different from the beta-binomial in several instances as well as for small and moderate sized samples. On the average, the standard error of Huynh's estimate was 50.2% and 56.9% of the standard error of the dual administration coefficient, respectively.

The bias of Subkoviak's kappa has not been investigated, and no studies have compared the bias and standard error of Subkoviak's and Huynh's kappa estimates. The same issues raised under the discussion of the bias of the $\underline{p}_o$ estimates are also relevant for kappa formulations. Specifically, these coefficients' biases and standard errors need to be evaluated for various score distributions, including a bimodal, and for situations where the classic parallelism assumption is violated.

Obviously, the lack of empirical research does not allow definitive recommendations as to which coefficient to use within each cell of Figure 2 given a particular situation. In order to address

some of the uninvestigated issues raised in this discussion, the current study was conducted to assess the influence of various test characteristics upon the bias and standard error associated with each major single test administration coefficient when estimating the appropriate dual test administration population coefficient. Specifically, the effects of the following variables were examined:

(1) violation of the classic parallelism assumption

(2) shape of the test score distribution

(3) test length

(4) cut-off score

(5) number of examinees in the sample

Those coefficients whose derivation is based upon the assumption of classically parallel tests were expected to be more biased when this assumption was violated (i.e., when the tests were randomly parallel). The shape of the test score distribution (particularly a bimodal distribution) was hypothesized to influence the bias of the threshold agreement indices because of their implicit or explicit distributional assumptions. The location of the cut-off was not expected to affect the extent of bias. Finally, a decrease in standard error was predicted as test length and sample size increased.

METHOD

Data Base

Several populations reflecting different distributional shapes were generated from data obtained from one of two sources. The first data base came from the responses of a sample of Michigan public school fourth graders to various criterion-referenced tests administered by the Michigan Educational Assessment Program (MEAP). MEAP annually collects data on fourth, seventh, and tenth grade students' attainment of various reading and mathematics objectives which address several of the minimal skills beginning students in these grades should have. Using a replicated, systematic sampling procedure, MEAP annually selects approximately 5000 students in each grade and computes each test's technical characteristics from their data (Michigan Department of Education, 1977). (In applying this sampling plan, the Michigan Department of Education (1977) randomly chooses ten numbers identifying the first member of each of ten systematic samples. A spacing factor is computed and added to each of these numbers to identify the next member of each set. The spacing factor is repeatedly added to the previous set of numbers until the requisite sample size has been attained.) The data obtained from a sampling of 5,040 fourth grade students in the fall of 1979 served as the major population data base in this study. The second data source or population was the responses of 589 college students to a mid-term exam given in their introductory psychology course. This exam was a "norm-referenced test" and produced a distribution not commonly found with criterion-referenced tests.

Test Ch

Di

the stu

with a

zero; a

that fo

tional

distrib

was inc

data.

is also

lower m

that a

questio

bution

shaped

some ca

bimodal

bimodal

Finally

priate

tests.

members

Procedure

Test Characteristics

Distribution shape. Four distributions were incorporated into the study: (1) severely negatively skewed; (2) J-shaped; (3) bimodal with a bigger mode at the upper end and a lower mode not equal to zero; and (4) normal. The first distribution was believed to typify that found when a criterion-referenced test is given after an instructional or a training program (Marshall, 1976). Correspondingly, this distribution was found in the MEAP data. The J-shaped distribution was included in the study because it was also represented in the MEAP data. According to Marshall and Serlin (1979), a bimodal distribution is also frequently found in mastery testing situations. Setting the lower mode unequal to zero was intended to reflect the probability that a non-master would guess the correct answer to one or more questions. Marshall and Serlin (1979) contended that this distribution is much more likely to occur in mastery testing than a J- or U-shaped distribution, especially when guessing is a viable factor. In some cases, the MEAP data (considering each grade) did follow a bimodal distribution with the lower mode equal to one. However, the bimodal did not occur more often than the J-shaped distribution. Finally, a normal distribution was included to explore the appropriateness of the reliability formulas for typical norm-referenced tests. Note that the bimodal and the normal distributions are not members of the beta distribution family.

I

examin

refere

a high

state

Novick

betwee

and be

decisi

C

employ

measur

1976;

and, c

(1972)

percen

60 and

seems

Instru

C

of tes

equall

fore,

length

percen

Test length. Test lengths of 5, 10, 15, and 20 items were examined. These test lengths typify those found for criterion-referenced tests and/or are representative of those needed to produce a high probability of accurately assigning respondents to a mastery state (Algina & Noe, 1978; Klein & Kosecoff, 1973; Marshall, 1976; Novick and Lewis, 1974). Furthermore, Berk (1980) recommended using between five and ten items per objective for most classroom decisions and between 10 and 20 items for school, system, and state level decisions.

Cut-off score. Three cut-off scores, 70%, 80%, and 90% were employed because they are representative of those occuring in mastery measurement and/or those recommended for usage (Block, 1972; Marshall, 1976; Novick & Lewis, 1974). To adequately effect cognitive learning and, concurrently, maintain interest in learning, Block's research (1972) has shown that the cut-off should be set between 80 and 85 percent. Marshall (1976) stated that one would typically use between 60 and 90 percent, and Novick and Lewis (1974) noted that the range seems to be between 70 and 85 percent in Individually Prescribed Instruction.

Given the previously specified test lengths and the integer value of test scores, specifying three test scores (advancement scores) equalling the chosen cut-off levels was not always possible. Therefore, reliabilities were only computed for those combinations of test length and cut-off score for which a test score resulting in a percentage equal to or slightly greater than the given cut-off could

be specif

ment scor

Numb

randomly

chosen be

illustra

organiza

because

formulas

for long

were bel

size.

Tes

Data Gen

Ite

randomly

drawn.

was need

included

be specified. Figure 3 presents these combinations and each advancement score with its associated cut-off level.

Number of examinees. Sample sizes of 25, 35, and 50 were randomly selected from the population. The first two values were chosen because they were believed to typify classroom sizes and to be illustrative of the number of people participating in various organizational training programs. A sample size of 50 was used because it has been recommended that estimation of α and β in Huynh's formulas be accomplished with $N > 40$ for very short tests and $N \geq 2n$ for longer tests (Subkoviak, 1978). Finally, these three sample sizes were believed to be divergent enough to study the effects of sample size.

Test Length	Cut-off Level		
	70%	80%	90%
5		4/5 (80%)	
10	7/10 (70%)	8/10 (80%)	9/10 (90%)
15	11/15 (73%)	12/15 (80%)	14/15 (93%)
20	14/20 (70%)	16/20 (80%)	18/20 (90%)

Figure 3.--Advancement Scores for Each Combination of Test Length and Cut-off Level

Data Generation

Item Domain. The study required a domain of items from which randomly and classically parallel tests of various lengths could be drawn. Specifically, a content domain consisting of at least 40 items was needed to construct alternate forms of all possible test lengths included in this study. Since all the MEAP criterion-referenced tests

consisted of five items, items had to be taken from at least eight tests, measuring different objectives, to form the domain. MEAP groups the mathematic and reading objectives into major skill areas. For example, the program includes 15 mathematics objectives tapping various aspects of numeration skill. A content analysis indicated that eight tests from the numeration skill area appeared to measure similar objectives. These 40 items were intercorrelated and subjected to a principal components analysis. The mean item intercorrelation within objectives was .36. The mean intercorrelation between items on different objectives, computed by systematically sampling correlations within the 40 x 40 correlation matrix, was .16. The principal components analysis yielded a general factor accounting for 21.4% of the variance. Ten factors had eigenvalues greater than or equal to one. A varimax rotation indicated that, generally, items within a particular test loaded highest on the same factor and each factor was defined by the items on one particular test. In summary, the set of 40 items was more heterogeneous than what one might find for a very narrowly defined objective. However, the KR-20 was .89, indicating a fairly high internal consistency. Therefore, the researcher decided to use these items to construct the domain.

Forty students did not reach the questions in one or more of the eight MEAP tests comprising the domain and were, therefore, eliminated from the data base. Based upon 5,000 students, the p values of the 40 items ranged from .69 to .96. The mean and standard deviation of domain scores were 35.74 and 5.34, respectively.

Th

46 item

items w

KR-20 w

standar

S

this st

bution

bimoda

distri

S

eight

surpri

item d

of thi

I

and me

closel

sampli

the wa

of the

closel

R

random

random

of the

The second data base, the psychology mid-term exam, consisted of 46 items. To increase this item domain's internal consistency, six items with low item-total correlations were eliminated. The resultant KR-20 was .68. The p values ranged from .19 to .96, and the mean and standard deviation of domain scores were 27.74 and 4.54, respectively.

Score Distributions. The reason for using two data sources in this study was to provide a population representative of each distribution under investigation. The negatively skewed, J-shaped, and bimodal distributions were based upon the MEAP data, while the normal distribution was represented by the psychology mid-term domain scores.

Similar to the majority of MEAP's criterion-referenced tests, the eight numeration tests produced negatively skewed distributions. Not surprisingly, the frequency distribution of total scores on the 40-item domain was also negatively skewed. Figure 4 presents the graph of this population distribution.

To generate the J distribution, the domain scores were inverted and merged with the original scores. The resulting distribution closely resembled a U. Then, a new population was formed by randomly sampling 3,500 students from the original distribution (upper half of the "U") and 1,500 students from the inverted distribution (lower half of the "U"). As can be seen in Figure 5, the graph of this population closely follows a J-shape.

The bimodal distribution was formed by altering the scores of a random sample of people from the negatively skewed distribution on a random sample of items. Specifically, the researcher first sampled 6% of those with scores greater than or equal to 30 and changed their

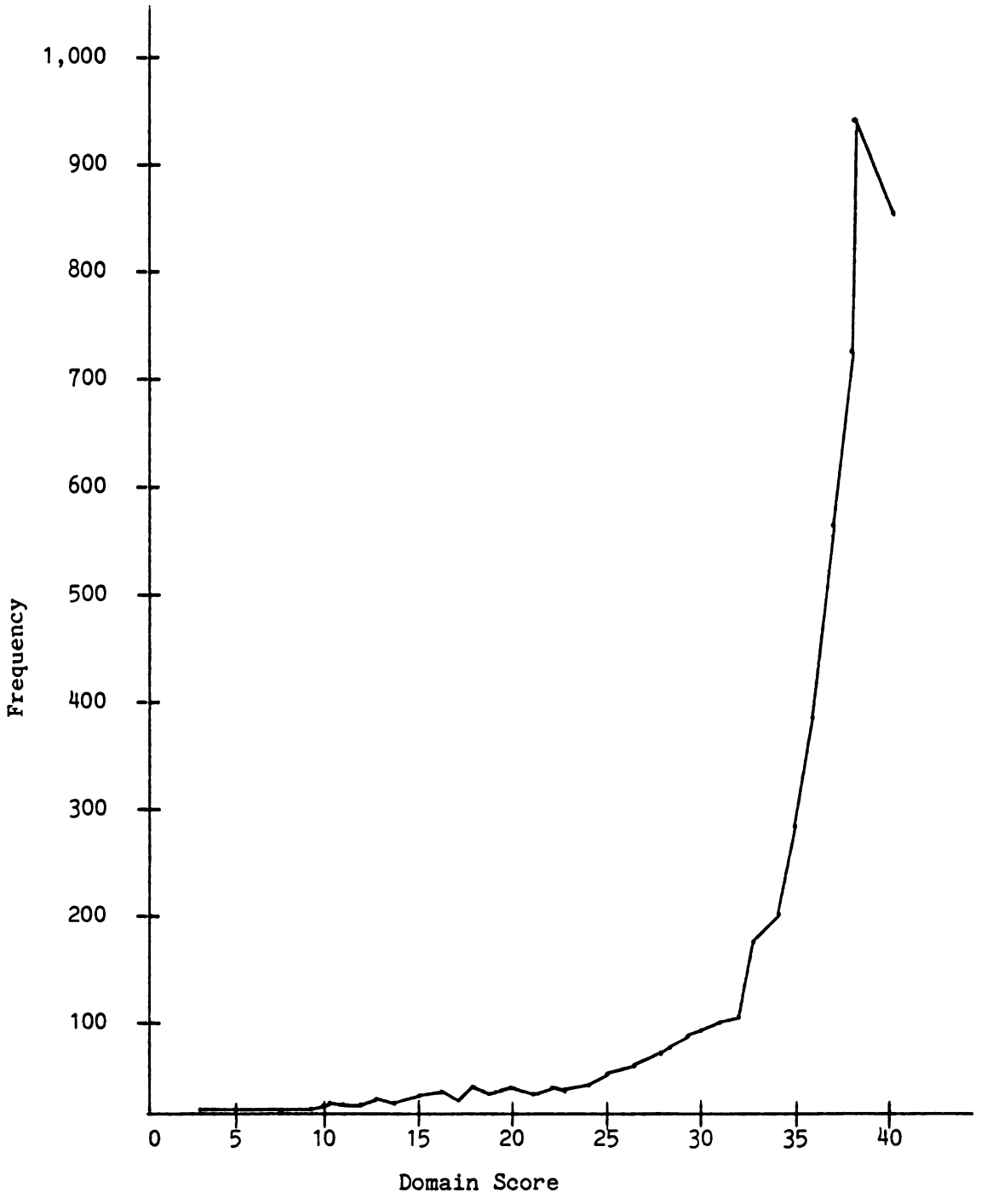


Figure 4.--Skewed Population Frequency Distribution of Domain Scores

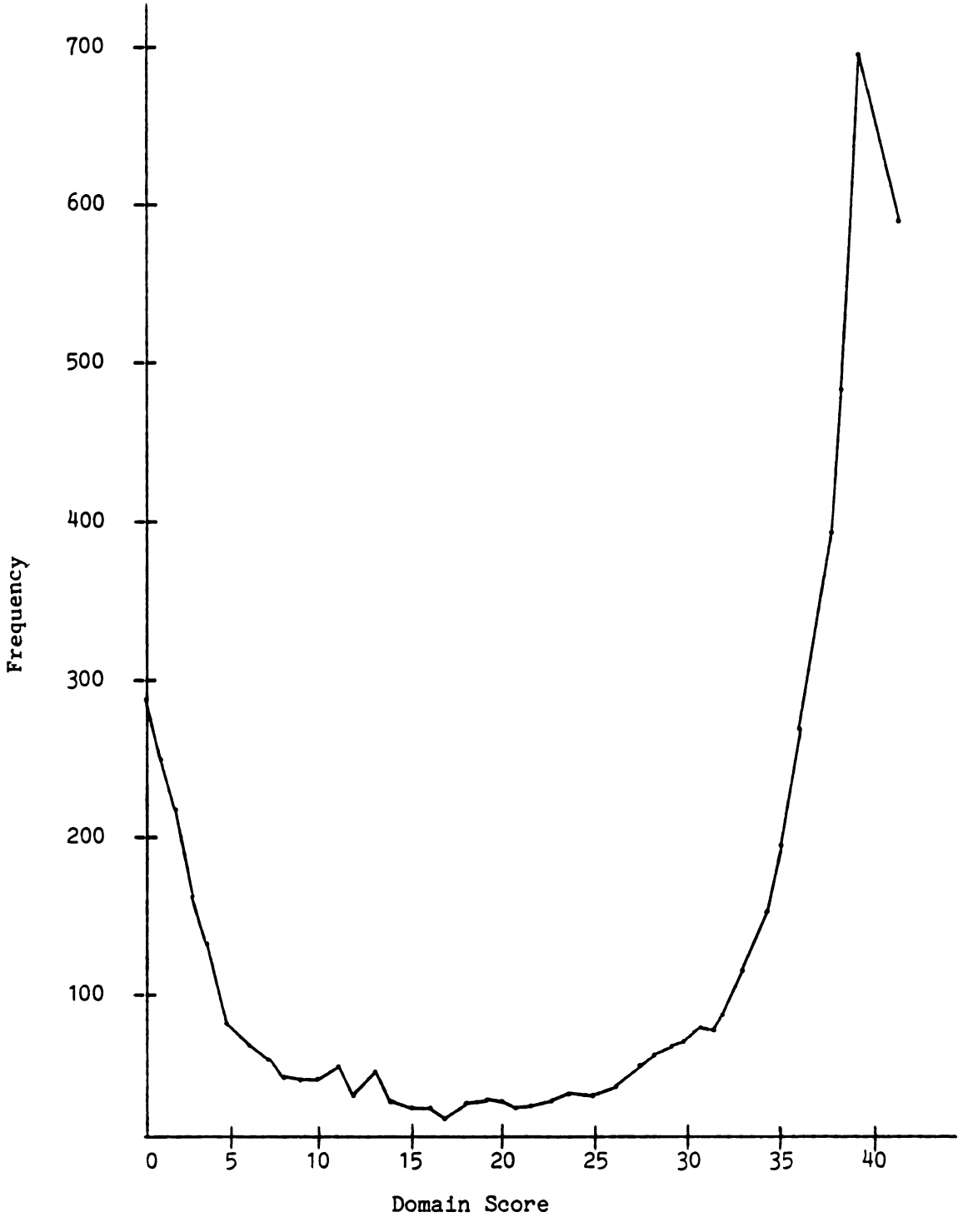


Figure 5.--J-shaped Population Frequency Distribution of Domain Scores

scores from right to wrong on a sample of 30 items. If a student had already answered a particular question wrong, the item response was not altered. The reason for changing 30 items was to assure that the lower mode would equal the number of items expected to be answered correctly merely by guessing. This same procedure was repeated two more times with replacement of items and people occurring between each sampling procedure. If a student was selected in more than one sampling procedure, he/she was deleted from the second and/or third sample. These three samples were combined with the unaltered scores in the original distribution, producing the population frequency distribution depicted in Figure 6.

Finally, the psychology mid-term scores were duplicated five times to create enough examinees for the sampling process. The resultant domain scores of 2,945 examinees produced the approximately normal distribution shown in Figure 7. The skewness and kurtosis moments were -.30 and .15, respectively. (In the computer package used in this study, the kurtosis of a normal distribution was zero instead of three.) These statistics indicated that the distribution was slightly negatively skewed and somewhat more peaked than a normal distribution. However, the departure did not appear to be practically significant.

Alternate forms. Following the construction of an item domain and the distribution manipulations, alternate parallel and randomly parallel forms were constructed for each test length. Randomly parallel five item tests were formed by randomly sampling items from the domain without replacement. Consequently, alternate test forms

800

700

600

500

Frequency

400

300

200

100

Figure

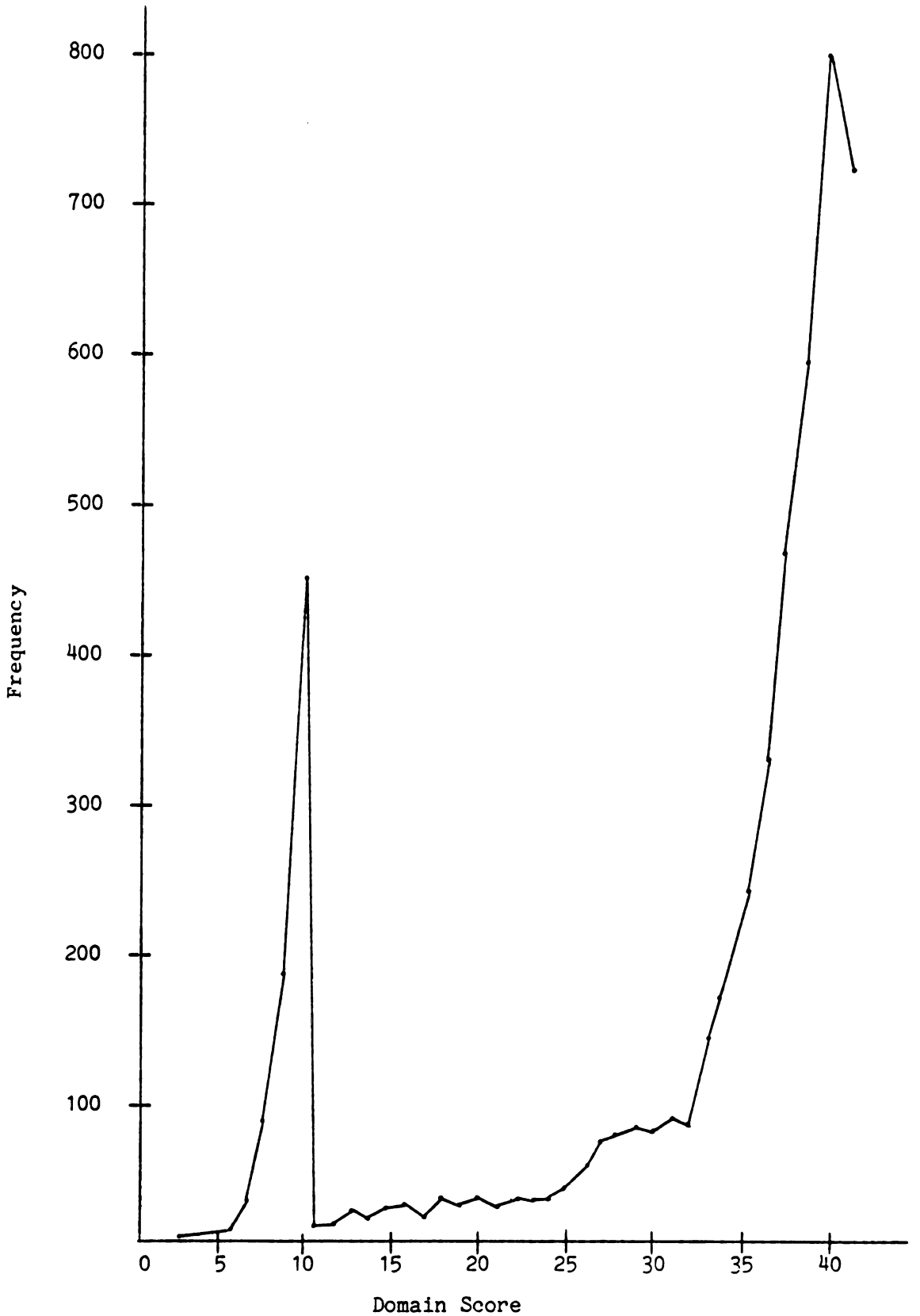


Figure 6.--Bimodal Population Frequency Distriution of Domain Scores

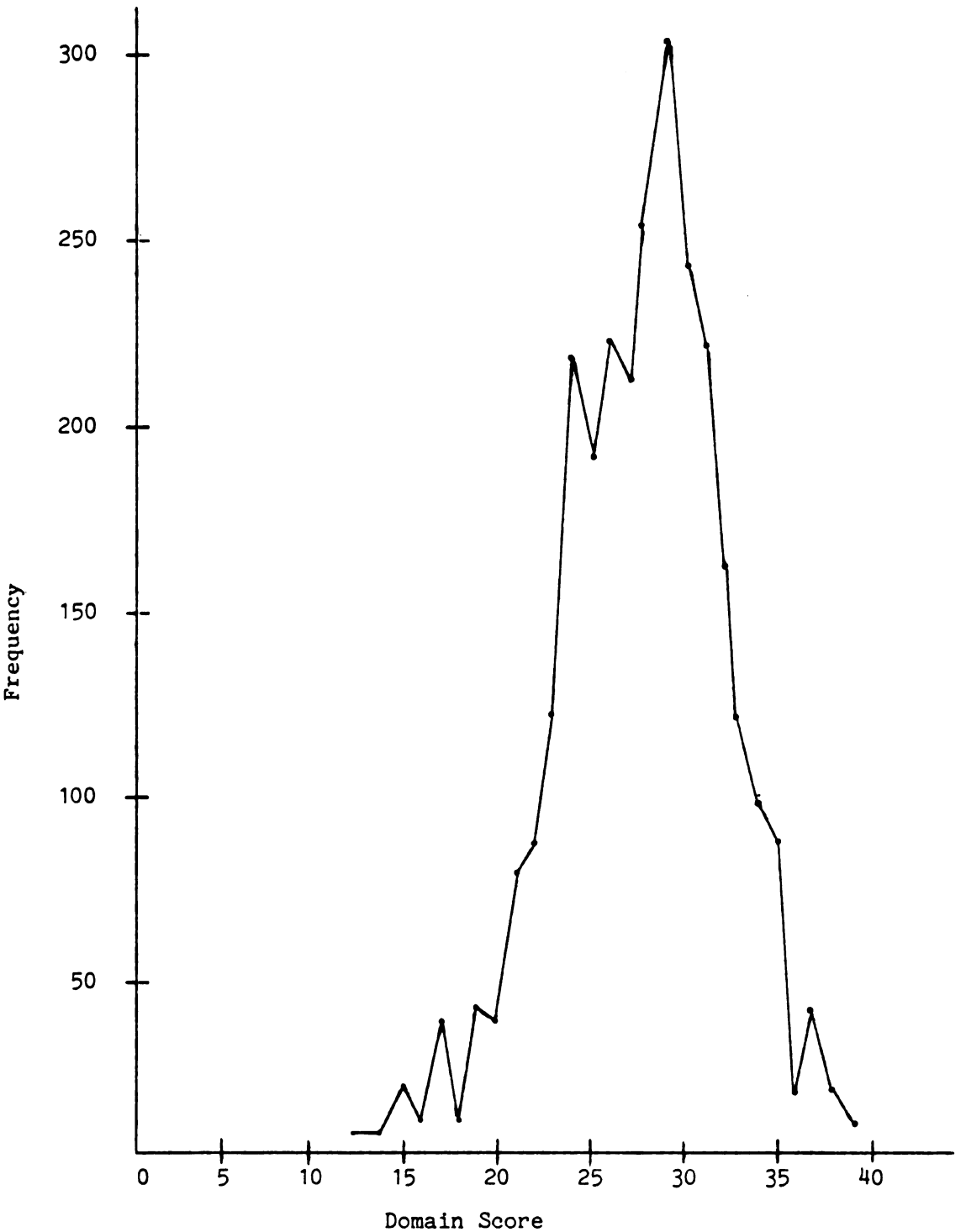


Figure 7.--Normal Population Frequency Distribution of Domain Scores

did

repl

alte

thos

(wit

thos

used

part

stru

samp

item

from

most

five

chos

valu

alte

dist

Dete

shap

Livi

Brep

exce

all

did not have any items in common. The items from both forms were not replaced in the domain when longer tests were constructed. For each alternate form, tests of 10, 15, and 20 items were built by using those items found on the next shorter test and randomly sampling (without replacement) the necessary number of additional items from those remaining in the domain. For the MEAP data, the same tests were used for the skewed and J distributions. However, since these particular tests did not produce bimodal distributions, the test construction process was repeated for the bimodal score domain. The sampling procedure was also repeated for the psychology exam data.

Alternate classically parallel forms were constructed by pairing items based on their $\underline{p}$ values and item-total correlations. One item from each pair was placed in each form. The five pairs having the most equivalent items within each pair were used to construct the five-item tests. In forming longer tests, the next closest pairs were chosen and added to those on the next shorter test. Since the $\underline{p}$ values and/or the item-total correlations were expected to change when altering the distribution shape, this process was repeated for each distribution.

Determination of Bias

For every combination of test length, cut-off score, distribution shape, and type of parallelism, population values of $\underline{p}_o$, kappa, and Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ were computed from two test administrations. Brennan and Kane's $\Phi(\lambda)$ and Φ were also computed in every condition except those involving classically parallel alternate forms because all items in the domain would have to have equal $\underline{p}$ values to meet this

assumption. Moreover, if all items had equal p values, $\Phi(\lambda)$ would simply equal Livingston's $K^2(X,T_X)$ and Φ would equal the generalizability coefficient (Brennan, 1978, 1979). (Note also that the value of Φ does not change as the cut-off score is altered.) In all, 320 population values were computed. Formulas for each population coefficient can be found in Figure 8.

Thirty independent random samples of 25, 35, and 50 cases were drawn with replacement from each of the four population distributions. Within each cell of the design, an estimate of each population coefficient was computed for each of the 30 samples using the appropriate single test administration coefficients. These estimates were obtained for only one alternate form. The mean of the estimates within each cell was compared to the population value to determine the magnitude and direction of bias. The standard deviation of these estimates indicated each coefficient's sampling error. The total design contained 240 cells (four distribution shapes, four test lengths, three cut-off scores for tests of 10, 15, and 20 items, one cut-off score for a five-item test, three sample sizes, and either classically or randomly parallel alternate forms).

One problem was encountered in sampling examinees; KR-20 and KR-21 for some samples were negative or equal to zero. Although negative reliability coefficients can be equated to zero and many of the coefficients can be computed when KR-20 equals zero, Huynh's p_o and kappa estimates cannot. For each case in which KR-20 or KR-21 was negative or zero, the random sampling process was repeated until other samples with positive coefficients were found.

Coefficient	Alternate Form Population Formulas	Sample Estimate Formulas
Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{x}})$	$\dfrac{\sigma_{(\underline{x},\underline{y})} + \left[(\mu_{\underline{x}}-\underline{C}_{\underline{x}})\ (\mu_{\underline{y}}-\underline{C}_{\underline{y}})\right]}{\sqrt{\left[\sigma_{\underline{x}}^2+(\mu_{\underline{x}}-\underline{C}_{\underline{x}})^2\right]\ \left[\sigma_{\underline{y}}^2+(\mu_{\underline{y}}-\underline{C}_{\underline{y}})^2\right]}}$	$\dfrac{(\alpha_{20}(\underline{s}_{\underline{x}}^2))+ (\bar{\underline{X}}-\underline{C}_{\underline{x}})^2}{\underline{s}_{\underline{x}}^2 + (\bar{\underline{X}}-\underline{C}_{\underline{x}})^2}$
Brennan and Kane's $\Phi(\lambda)$	$\dfrac{\sigma_{\underline{p}}^2 + (\mu-\lambda)^2}{\sigma_{\underline{p}}^2 + (\mu-\lambda)^2 +(\sigma_{\underline{i}}^2/\underline{n}_{\underline{i}})+(\sigma_{\underline{p}\underline{i}}^2/\underline{n}_{\underline{i}})}$	$\dfrac{\underline{s}_{\underline{p}}^2+(\underline{s}_{\underline{p}\underline{i}}^2/40)+(\underline{X}_{\underline{P}\underline{I}}-\lambda)^2-\left\{(\underline{s}_{\underline{p}}^2/\underline{n}_{\underline{p}}) + \left[(1-(\underline{n}_{\underline{i}}/40))\ (\underline{s}_{\underline{i}}^2/\underline{n}_{\underline{i}})\right]+ (\underline{s}_{\underline{p}\underline{i}}^2/\underline{n}_{\underline{p}}\underline{n}_{\underline{i}})\right\}}{\underline{s}_{\underline{p}}^2+(\underline{s}_{\underline{p}\underline{i}}^2/40)+(\underline{X}_{\underline{P}\underline{I}}-\lambda)^2-\left\{(\underline{s}_{\underline{p}}^2/\underline{n}_{\underline{p}}) + \left[(1-(\underline{n}_{\underline{i}}/40))\ (\underline{s}_{\underline{i}}^2/\underline{n}_{\underline{i}})\right]+ (\underline{s}_{\underline{p}\underline{i}}^2/\underline{n}_{\underline{p}}\underline{n}_{\underline{i}})\right\} + \left\{\left[1-(\underline{n}_{\underline{i}}/40)\right]\ \left[(\underline{s}_{\underline{i}}^2+\underline{s}_{\underline{p}\underline{i}}^2)/\underline{n}_{\underline{i}}\right]\right\}}$
Brennan and Kane's Φ	$\dfrac{\sigma_{\underline{p}}^2}{\sigma_{\underline{p}}^2 +(\sigma_{\underline{i}}^2/\underline{n}_{\underline{i}})+(\sigma_{\underline{p}\underline{i}}^2/\underline{n}_{\underline{i}})}$	$\dfrac{\underline{s}_{\underline{p}}^2 + (\underline{s}_{\underline{p}\underline{i}}^2/40)}{\underline{s}_{\underline{p}}^2 +(\underline{s}_{\underline{p}\underline{i}}^2/40) +\left\{\left[1-(\underline{n}_{\underline{i}}/40)\right]\ \left[(\underline{s}_{\underline{i}}^2+\underline{s}_{\underline{p}\underline{i}}^2)/\underline{n}_{\underline{i}}\right]\right\}}$
Subkoviak's $\underline{P}_{\underline{o}}$	$\sum_{\underline{k}=1}^{\underline{m}} \underline{P}_{\underline{k}\underline{k}}$	$\dfrac{1}{\underline{n}_{\underline{p}}} \sum_{\underline{X}=0}^{\underline{n}_{\underline{i}}} \underline{f}(\underline{x}) \cdot \left[\underline{P}(\underline{X} \geq \underline{C}_{\underline{x}})^2 + (1-\underline{P}(\underline{X}>\underline{C}_{\underline{x}}))^2\right]$ where $\underline{P}(\underline{X}\geq\underline{C}_{\underline{x}})=\sum_{\underline{j}=\underline{C}_{\underline{x}}}^{\underline{n}_{\underline{i}}} \binom{\underline{n}_{\underline{i}}}{\underline{j}} \hat{\theta}_{\underline{x}}^{\underline{j}}\ (1-\hat{\theta}_{\underline{x}})^{\underline{n}_{\underline{i}}-\underline{j}}$ and $\hat{\theta}_{\underline{x}} = \left[\hat{\alpha}_{20}\left(\dfrac{\underline{X}}{\underline{n}_{\underline{i}}}\right)\right]+\left[(1-\hat{\alpha}_{20})\left(\dfrac{\bar{\underline{X}}}{\underline{n}_{\underline{i}}}\right)\right]$

Figure 8.--Formulas for Both Criterion-Referenced Reliability Population Coefficients Computed from Alternate Forms and Single Test Administration Sample Estimates

Coefficient	Alternate Form Population Formulas	Sample Estimate Formulas
Marshall and Haertel's p_o	$\sum_{k=1}^{m} p_{kk}$	$\frac{1}{n_p}\left[\sum_{y=0}^{C_x-1} \hat{f}(y) + \sum_{y=C_x}^{2C_x-2} \hat{f}(y) \cdot H_y(y-(C_x-1), C_x-1) + \sum_{y=2C_x}^{n_i+C_x-1} \hat{f}(y) \cdot H_y(C_x, y-C_x) + \sum_{y=n_i+C_x}^{2n_i} \hat{f}(y)\right]$ where $H_y(1,m) = \sum_{j=1}^{m}\left(\frac{\binom{y}{j}\binom{2n_i-y}{n_i-j}}{\binom{2n_i}{n_i}}\right)$; $\hat{f}(y) = \sum_{x=0}^{n_i} f(x)\binom{2n_i}{y}\cdot \hat{\theta}_x^{\,y}(1-\hat{\theta})_x^{2n_i-y}$; $\hat{\theta}_x = \left[\hat{\alpha}_{20}\left(\frac{\bar{x}}{n_i}\right)\right] + \left[(1-\hat{\alpha}_{20})\left(\frac{\bar{x}}{n_i}\right)\right]$
Huynh's p_o	$\sum_{k=1}^{m} p_{kk}$	See Huynh (1976)
Subkoviak's kappa	$\frac{\sum_{k=1}^{m} p_{kk} - \sum_{k=1}^{m} p_{k.}p_{.k}}{1 - \sum_{k=1}^{m} p_{k.}p_{.k}}$	$\frac{p_o - \left\{1-2\left[\sum_{j=1}^{n_p} P(X_j \geq C_x)/n_p) - (\sum_{j=1}^{n_p} P(X_j \geq C_x)/n_p)^2\right]\right\}}{1 - \left\{1-2\left[\sum_{j=1}^{n_p} P(X_j \geq C_x)/n_p) - (\sum_{j=1}^{n_p} P(X_j \geq C_x)/n_p)^2\right]\right\}}$
Huynh's kappa	$\frac{\sum_{k=1}^{m} p_{kk} - \sum_{k=1}^{m} p_{k.}p_{.k}}{1 - \sum_{k=1}^{m} p_{k.}p_{.k}}$	See Huynh (1976)

Figure 8 (cont'd.)

Note:

C_y or C_x	Cut-off score expressed as the number of items answered correctly.
λ	Cut-off score expressed as the proportion of items answered correctly.
μ	Grand mean in the population of persons and the domain or universe of items.
σ^2_p, s^2_p	Variance over persons of their universe scores.
σ^2_i, s^2_i	Variance of the item mean scores.
σ^2_{pi}, s^2_{pi}	Variance due to the interaction of persons and items plus experimental error.
X_{pI}	Mean over a random sample of persons and items.
n_p	Number of persons in the sample.
n_i	Number of test items.
$p_{k.}$, $p_{.k}$	Proportion of people placed in mastery state k on one test administration.
p_{kk}	Proportion of people consistently placed in mastery state k in both test administrations.

Within each distribution, the same samples were used to compute estimates of the coefficients for every combination of test length and cut-off score. Furthermore, the same samples were used for estimating the reliability of randomly and classically parallel tests. As mentioned previously, when a test had a zero or a negative KR-20 or KR-21 in a particular sample, the sample was eliminated and another one was chosen. However, only the internal consistency of five-item tests comprised of randomly chosen items was examined in determining which samples to delete. Since the classically parallel forms consisted of different items and since the same set of samples was used in both parallelism conditions, some samples retained in the set had a negative or zero KR-21 for the classically parallel form. This problem occurred only for the normal distribution and was probably due to the relatively low internal consistency of the items in this domain. Moreover, within this distribution, the KR-21 for longer tests within both parallelism conditions was negative or zero for some of the retained samples. In those cells where this difficulty surfaced, the sample(s) was dropped from the cell. Therefore, within some cells, the mean and standard deviation were based on less than 30 samples. However, every cell contained at least 20 samples.

Estimation Formulas. Figure 8 presents the single test administration formulas used to estimate each population alternate form coefficient. A few formulas require some explanation. In estimating $\Phi(\lambda)$, Brennan and Kane (1977a) noted that $(\bar{X} - \lambda)^2$ is not an unbiased

estimate of $(\mu - \lambda)^2$. They presented an unbiased estimate of this term:

$$(\underline{X}_{pI} - \lambda)^2 - \left(\frac{\underline{s}_{\underline{p}}^2}{\underline{n}_{\underline{p}}} + \frac{\underline{s}_{\underline{i}}^2}{\underline{n}_{\underline{i}}} + \frac{\underline{s}_{\underline{pi}}^2}{\underline{n}_{\underline{p}}\underline{n}_{\underline{i}}} \right)$$

In addition, previous discussion of Brennan and Kane's indices assumed the item domain was infinite. However, the domain in this study is a finite universe. To account for this design factor, Brennan (1978) provided formulas for $\Phi(\lambda)$ and Φ in which a finite universe correction factor is applied to the variance components comprising these coefficients. These latter formulas which also incorporate an unbiased estimate of $(\mu - \lambda)^2$ were used in this study and appear in Figure 8.

For Huynh's, Subkoviak's, and Marshall's indices, the researcher assumed that an individual's test scores followed a binomial distribution given his/her true score, rather than a compound binomial model. Studies cited previously have indicated that using the binomial model for heterogeneous item difficulty values does not substantially affect the accuracy of these coefficients. Moreover, the binomial model has produced results similar to those found using the compound binomial for Subkoviak's $\hat{\underline{p}}_{\underline{o}}$ (Marshall & Serlin, 1979).

For Marshall's $\hat{\underline{p}}_{\underline{o}}$, scores on a $2\underline{n}$-item test were simulated via a binomial regression model. Specifically, a linear regression was used to predict true score from obtained score and the predicted true score was used in a binomial error model to estimate the frequency distribution of a $2\underline{n}$-item test. As noted previously, Marshall and Serlin (1979) used five different models for simulating scores. The binomial regression model was chosen over the others because the relative size

of Marshall's $\hat{p}_{Q}$ using this model better reflected the distance between the cut-off and the mode(s) for distributions similar to those used herein.

RESULTS

Population Values

Tables 1 and 2 present the population distributional characteristics associated with each randomly and classically parallel alternate form, respectively. For the bimodal distribution within the randomly parallel condition, one five-item form had only one mode and one ten-item form had three modes. As can be seen from the skewness and kurtosis moments, the normal distributions departed from their theoretical shape.

For each condition, Tables 3 to 8 present the alternate form population values of the classical reliability coefficient (ρ_{11}), Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, Brennan and Kane's $\Phi(\lambda)$, Brennan and Kane's Φ, $\underline{p}_{\underline{o}}$, and kappa. To compute the kappa coefficient for classically parallel tests, the average of the corresponding marginal probabilities was used to determine the probability of chance agreement. Similarly, for $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, the average of the classically parallel tests' means and variances were used as the values of $\mu_{\underline{x}}$ ($\mu_{\underline{y}}$) and $\sigma^2_{\underline{x}}$ ($\sigma^2_{\underline{y}}$), respectively.

As can be seen in Table 3, ρ_{11} increased as test length increased. In general, given a particular distribution and test length, ρ_{11} was higher in the classically parallel condition than in the randomly parallel condition. The exceptions occurred for shorter tests. Comparing the results for each distribution, it becomes clear that tests derived from more internally consistent domains had higher alternate form reliabilities than those from domains in which the item intercorrelations were not as high.

Table 1.--Characteristics of Each Randomly Parallel Alternate Form.

Distribution	Test Length	Mean	Mode	Standard Deviation	Skewness	Kurtosis
Skewed	5	4.55 / 4.62	5 / 5	.85 / .75	-2.41 / -2.49	6.38 / 7.13
	10	8.83 / 9.21	10 / 10	1.67 / 1.36	-1.86 / -2.56	3.45 / 7.77
	15	13.29 / 13.53	15 / 15	2.36 / 2.07	-2.03 / -2.22	4.40 / 5.62
	20	17.63 / 18.11	20 / 20	3.09 / 2.58	-2.02 / -2.33	4.33 / 6.30
J-Shaped	5	3.33 / 3.35	0,5 / 0,5	2.06 / 2.08	-.73 / -.76	-1.22 / -1.21
	10	6.55 / 6.68	0,10 / 0,10	3.88 / 4.09	-.71 / -.77	-1.18 / -1.21
	15	9.84 / 9.92	0,15 / 0,15	5.79 / 5.90	-.73 / -.76	-1.18 / -1.19
	20	13.08 / 13.25	0,20 / 0,20	7.63 / 7.86	-.73 / -.77	-1.18 / -1.19

Table 1 (cont'd.)

Distribution	Test Length	Mean	Mode	Standard Deviation	Skewness	Kurtosis
Bimodal	5	3.71 / 4.13	5 / 1,5	1.54 / 1.39	-1.10 / -1.41	.11 / .46
	10	7.74 / 8.05	1,3,10 / 3,10	2.76 / 2.74	-1.21 / -1.29	.22 / .14
	15	11.57 / 12.02	4,15 / 4,15	4.14 / 4.12	-1.16 / -1.34	-.11 / .25
	20	15.41 / 16.20	5,20 / 5,20	5.50 / 5.30	-1.20 / -1.38	-.05 / .32
Normal	5	3.71 / 2.71	4 / 3	1.09 / 1.13	-.70 / .02	.18 / -.46
	10	7.09 / 5.42	8 / 6	1.63 / 1.67	-.45 / .02	-.01 / -.42
	15	10.99 / 9.26	12 / 9	2.10 / 2.08	-.48 / -.08	.14 / -.27
	20	14.46 / 13.28	14 / 14	2.63 / 2.64	-.40 / -.29	.04 / -.01

Table 2.--Characteristics of Each Classically Parallel Alternate Form.

Distribution	Test Length	Mean	Mode	Standard Deviation	Skewness	Kurtosis
Skewed	5	4.60 / 4.58	5 / 5	.80 / .82	-2.49 / -2.45	6.89 / 6.89
	10	9.11 / 9.11	10 / 10	1.44 / 1.37	-2.32 / -2.31	6.30 / 6.49
	15	13.67 / 13.64	15 / 15	1.95 / 1.94	-2.50 / -2.36	7.54 / 6.58
	20	17.84 / 17.90	20 / 20	2.78 / 2.79	-2.20 / -2.16	5.61 / 5.10
J-Shaped	5	3.31 / 3.30	0,5 / 0,5	2.02 / 2.01	-.72 / -.72	-1.20 / -1.19
	10	6.69 / 6.69	0,10 / 0,10	4.10 / 4.07	-.77 / -.78	-1.21 / -1.20
	15	9.95 / 9.95	0,15 / 0,15	5.97 / 5.93	-.77 / -.77	-1.20 / -1.20
	20	13.15 / 13.18	0,20 / 0,20	7.71 / 7.73	-.76 / -.75	-1.19 / -1.18

Table 2 (cont'd)

Distribution	Test Length	Mean	Mode	Standard Deviation	Skewness	Kurtosis
Bimodal	5	4.09 / 4.11	1,5 / 2,5	1.28 / 1.25	-1.31 / -1.19	.58 / .14
	10	8.13 / 8.18	3,10 / 3,10	2.54 / 2.48	-1.25 / -1.31	.11 / .25
	15	11.96 / 11.99	4,15 / 4,15	4.09 / 4.05	-1.34 / -1.35	.25 / .29
	20	15.74 / 15.87	4,20 / 4,20	5.41 / 5.37	-1.29 / -1.32	.14 / .21
Normal	5	3.84 / 3.82	4 / 4	.98 / .97	-.60 / -.55	-.18 / -.17
	10	7.19 / 7.21	8 / 8	1.63 / 1.60	-.57 / -.40	.10 / -.06
	15	10.74 / 10.84	12 / 12	2.17 / 2.15	-.44 / -.46	-.12 / -.04
	20	13.88 / 13.86	15 / 15	2.63 / 2.58	-.33 / -.34	-.10 / .11

Table 3.--Classical Reliability of Randomly and Classically Parallel Alternate Forms for Each Distribution/Test Length Combination.

	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
Test Length	Skewed	J	Bi-modal	Normal	Skewed	J	Bi-modal	Normal
5	.60	.94	.77	.22	.66	.93	.80	.16
10	.62	.94	.89	.31	.69	.97	.90	.40
15	.74	.96	.93	.40	.77	.98	.94	.46
20	.77	.97	.95	.48	.83	.98	.95	.52

Table 4.--Alternate Form Population Values of Livingston's $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ for Each Cell of the Design.

		Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
Test Length	Cut-off Score	Skewed	J	Bi-modal	Normal	Skewed	J	Bi-modal	Normal
5	4	.736	.945	.734	.336	.774	.940	.800	.181
	7	.847	.939	.899	.190	.904	.975	.917	.413
10	8	.709	.945	.888	.558	.809	.977	.899	.523
	9	.591	.954	.905	.775	.692	.981	.909	.735
	11	.876	.962	.927	.309	.921	.977	.946	.466
15	12	.809	.965	.925	.563	.869	.979	.942	.590
	14	.756	.973	.942	.843	.782	.984	.954	.832
	14	.911	.966	.941	.416	.942	.980	.956	.518
20	16	.824	.969	.936	.654	.883	.982	.951	.710
	18	.761	.975	.947	.842	.830	.985	.958	.862

Table 5.--Population Values of Brennan and Kane's $\Phi(\lambda)$ for Each Cell of the Design.

		Randomly Parallel Alternate Forms			
Test Length	Cut-off Score	Skewed	J	Bi-modal	Normal
5	4	.632	.915	.789	.378
	7	.877	.951	.893	.393
10	8	.774	.956	.882	.548
	9	.697	.964	.897	.736
	11	.894	.967	.921	.521
15	12	.837	.970	.918	.645
	14	.790	.977	.935	.841
	14	.935	.975	.943	.564
20	16	.873	.977	.937	.708
	18	.822	.981	.946	.848

Table 6.--Population Values of Brennan and Kane's Φ for Each Cell of the Design.

		Randomly Parallel Alternate Forms			
Test Length	Cut-off Score	Skewed	J	Bi-modal	Normal
5	4	.535	.905	.789	.244
	7	.697	.950	.882	.392
10	8	.697	.950	.882	.392
	9	.697	.950	.882	.392
	11	.775	.966	.918	.492
15	12	.775	.966	.918	.492
	14	.775	.966	.918	.492
	14	.821	.974	.937	.563
20	16	.821	.974	.937	.563
	18	.821	.974	.937	.563

Table 7.--Alternate Form Population Values of p_o for Each Cell of the Design.

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J	Bi-modal	Normal	Skewed	J	Bi-modal	Normal
5	4	.927	.943	.852	.499	.935	.928	.925	.613
10	7	.931	.945	.938	.497	.957	.977	.966	.667
	8	.873	.908	.891	.621	.917	.964	.929	.657
	9	.783	.847	.828	.815	.826	.904	.829	.733
15	11	.926	.943	.931	.569	.963	.967	.953	.643
	12	.896	.925	.897	.615	.934	.950	.933	.650
	14	.760	.835	.782	.893	.766	.851	.791	.879
20	14	.940	.953	.944	.645	.956	.965	.953	.676
	16	.898	.929	.908	.683	.927	.948	.927	.749
	18	.798	.862	.833	.888	.830	.885	.859	.900

Table 8.--Alternate Form Population Values of Kappa for Each Cell of the Design.

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J	Bi-modal	Normal	Skewed	J	Bi-modal	Normal
5	4	.517	.875	.635	.106	.591	.846	.792	.135
10	7	.505	.879	.824	.132	.611	.948	.897	.221
	8	.435	.807	.735	.162	.572	.922	.809	.310
	9	.387	.692	.637	.120	.461	.803	.622	.214
15	11	.579	.878	.815	.215	.705	.928	.864	.270
	12	.579	.844	.752	.154	.636	.893	.824	.279
	14	.462	.668	.566	.107	.439	.703	.576	.305
20	14	.593	.896	.842	.296	.690	.924	.865	.335
	16	.589	.851	.777	.253	.696	.892	.819	.372
	18	.500	.725	.655	.233	.580	.771	.705	.283

Several characteristics of the criterion-referenced coefficients deserve attention. First, not surprisingly, those computed using classically parallel tests were generally greater than their counterparts in the randomly parallel condition. (This comparison was, of course, only relevant for $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, $\underline{p}_{\underline{o}}$, and kappa.) The exceptions appeared to be related to the size of ρ_{11} and the location of the cut-off. For example, within the J distribution, Table 3 indicates that ρ_{11} of the 5-item randomly parallel tests was slightly greater than its classically parallel counterpart. Likewise, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$, $\underline{p}_{\underline{o}}$, and kappa were also higher in the randomly parallel condition. In other cases, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ was higher in the randomly parallel condition even though ρ_{11} was lower. In these instances, the means of the randomly parallel tests were further from the cut-off than the means of the classically parallel tests. As Shavelson et al. (1972) noted, the difference between the cut-off and the mean can influence $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ more than ρ_{11} does. For $\underline{p}_{\underline{o}}$ and kappa, the relationship of the cut-off to heavy score density areas and to the size of the chance agreement probability appeared to account for the other exceptions.

Second, $\Phi(\lambda)$ and Φ increased as test length increased. Except for a few instances in the randomly parallel condition, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ was also an increasing function of test length. Contrary to previous findings, $\underline{p}_{\underline{o}}$ and kappa did not follow this trend even though ρ_{11} increased (Eignor & Hambleton, 1979; Subkoviak, 1978). This latter result indicates that the size of the error (expressed as a proportion) found in classical reliability may not correlate with the proportion of error found in reliability coefficients based on the Platonic true score model.

Third, given a particular test length, $\underline{p}_o$ increased as the cut-off moved away from heavy score density areas. For the skewed, J, and bimodal distributions, these areas were in the upper extremes of the distribution. Although $\underline{p}_o$ has been known to increase as the cut-off approaches the extremes, the score density appears to have had more influence on the size of $\underline{p}_o$ in this study. Except for the normal distribution, the changes in the value of kappa as a function of the cut-off generally followed the same pattern as $\underline{p}_o$. One might expect kappa to become higher as the cut-off approaches denser areas because the probability of chance agreement decreases. However, the author believes that due to the large size of these dense areas in the skewed, J, and bimodal distributions, $\underline{p}_o$ was reduced enough to outweigh this factor. For the normal distribution, the strength of the heavy score density areas and the size of the chance agreement probability also appeared to interact, producing some unusual patterns of kappa coefficients. Finally, as expected, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ increased as the distance between the cut-off and the mean increased.

Bias

Appendices A1 to A24 present the mean bias and standard deviation of each single test administration coefficient for each cell of the design. A negative value indicates underestimation, and a positive value means that the single test administration coefficient overestimated its population value. Except in two instances, the results for Subkoviak's and Marshall's $\underline{p}_o$ estimates were equal, confirming Marshall and Serlin's findings (1979). For the two exceptions, one for bias and one for standard deviation, the results differed by only .001, indicating that

the differences may simply be due to rounding error. Therefore, to avoid redundancy, only one of these coefficients, Subkoviak's $\hat{p}_o$, is mentioned and discussed below. The reader should assume that this discussion applies equally to Marshall's $\hat{p}_o$. To investigate each hypothesis, the mean of these statistics across appropriate cells was computed. In doing so, each cell's mean and standard deviation was weighted by the number of samples upon which it was based.

Throughout the ensuing discussion, the use of the term "significance" means practical significance, rather than statistical significance. For this study, any mean biases and standard deviations greater than or equal to .025 and differences between mean biases and standard deviations greater than or equal to this value were considered practically significant.

The relative ability of the coefficients to estimate their respective population reliability coefficients for randomly versus classically parallel tests was examined for $K^2(X,T_X)$, p_o, and kappa since their single test administration estimates assume classic parallelism. Collapsing across number of examinees, distribution type, test length, and cut-off score, Table 9 contains the mean bias of each estimate for both types of parallelism. Contrary to expectation, the absolute mean bias in the randomly parallel condition was less than or equal to that in the classically parallel condition for every coefficient. However, the only significant difference between the two conditions was for Subkoviak's $\hat{K}$. Taking direction into account, violation of the classic parallelism assumption significantly altered the mean bias of $\hat{K}^2(X,T_X)$ and the kappa estimates, while the p_o estimates were fairly robust. In

Table 9.--Mean Bias (Across Cells) of Various Coefficients in Estimating the Reliability of Classically and Randomly Parallel Alternate Forms.

	Type of Parallelism	
Coefficient	Random	Classic
Livingston's $\hat{K}^2(X,T_X)$	.019	-.019
Subkoviak's $\hat{p}_o$	-.010	-.030
Huynh's $\hat{p}_o$	.011	-.012
Subkoviak's $\hat{K}$	-.023	-.111
Huynh's $\hat{K}$	.039	-.050

the classically parallel case, all indices underestimated the population coefficient with the kappa estimates and Subkoviak's $\hat{p}_o$ doing so significantly. Given randomly parallel tests, only Subkoviak's coefficients were underestimates. The others overestimated their corresponding parameters with Huynh's $\hat{K}$ being a significant overestimate. In previous research, Huynh's coefficients have always been underestimates (Huynh & Saunders, 1979; Subkoviak, 1978). However, these studies used equivalent tests. The present findings support the past research, but also indicate that past results do not generalize to the randomly parallel condition.

The second hypothesis was that the p_o and kappa estimates would be more biased for those distributions not belonging to the beta-binomial family (i.e., bimodal and normal). Even though no hypotheses were generated for the influence of the distribution upon $K^2(X,T_X)$, $\Phi(\lambda)$, and Φ, Table 10 presents the mean bias of every coefficient for each distribution. Based upon the absolute value of the mean bias, the

Table 10.--Mean Bias Across Cells of Each Reliability Coefficient for Each Distribution.

	Distribution			
Coefficient	Skewed	J-Shaped	Bimodal	Normal
Livingston's $\hat{K}^2(X,T_X)$	.004	-.005	-.025	.028
Brennan & Kane's $\hat{\Phi}(\lambda)$[a]	.057	.008	.010	.061
Brennan & Kane's $\hat{\Phi}$[a]	.086	.009	.011	.141
Subkoviak's $\hat{p}_o$	-.020	-.011	-.028	-.019
Huynh's $\hat{p}_o$	-.011	.018	-.011	.000
Subkoviak's $\hat{K}$	-.076	-.032	-.069	-.092
Huynh's $\hat{K}$	-.018	.034	-.025	-.014

[a]Means for these coefficients were based only on cells within the randomly parallel condition.

pattern of results for Subkoviak's coefficients conformed somewhat to that predicted. Specifically, Subkoviak's $\hat{p}_o$ and $\hat{K}$ were least biased for the J distribution and most biased given the bimodal and the normal distributions, respectively. In the case of Subkoviak's $\hat{K}$, the differences between the J and the other distributions were significant. Contrary to expectation, Subkoviak's $\hat{p}_o$ was almost equally biased for the skewed and normal distributions, and Subkoviak's $\hat{K}$ was slightly more biased for the skewed than for the bimodal. Generally, the absolute mean bias of Huynh's coefficients followed a pattern opposite to that predicted; Huynh's $\hat{p}_o$ and $\hat{K}$ were least biased for the normal distribution and most biased for the J distribution. However, as expected, Huynh's $\hat{K}$ was less accurate for the bimodal than for the skewed

distribution. For Huynh's $\hat{p}_o$, the biases associated with these two distributions were equal and in the same direction. In no case did any distribution significantly change the absolute mean bias of Huynh's coefficients.

Considering both magnitude and direction, Subkoviak's $\hat{p}_o$ consistently underestimated the population value with significant bias occurring for the bimodal distribution. Note, however, that the mean bias of this coefficient was not significantly altered by changes in the distribution's shape, regardless of whether or not the type of distribution violated the underlying assumptions. On the other hand, altering the distribution changed the direction of bias for Huynh's $\hat{p}_o$, leading to significant differences between the results for the J distribution and those found for the skewed and bimodal distributions. Specifically, Huynh's $\hat{p}_o$ was unbiased for the normal distribution, slightly negatively biased for the skewed and bimodal, and positively biased for the J distribution. In no case were these degrees of bias significant. For both kappa estimates, the bias associated with the J distribution was significantly different from that found in the other conditions. Specifically, Subkoviak's $\hat{K}$ underestimated kappa much more for the other distributions, although the extent of bias was significant throughout. In the case of Huynh's $\hat{K}$, the J distribution significantly affected the direction of bias as it had done for Huynh's $\hat{p}_o$; Huynh's $\hat{K}$ was positively biased for the J distribution and negatively biased for the others. In addition, the biases associated with the J and bimodal distributions were significant.

Table 10 indicates that the bias of $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ was significantly affected by the type of distribution in terms of both magnitude and direction. The biases for the bimodal and normal distributions differed significantly with significant overestimation associated with the former and an approximately equal, but negative, bias corresponding to the latter distribution. The mean biases for the skewed and J distributions were close to zero and were significantly different from those found for the bimodal and normal distributions, respectively.

$\hat{\Phi}(\lambda)$ and $\hat{\Phi}$ followed the same pattern. They consistently overestimated their parameters and were significantly less accurate for the normal and skewed distributions than for the others. As a matter-of-fact, the extent of bias associated with the normal and skewed distributions was quite high and significant, but was very slight for the other distributions. For $\hat{\Phi}$, the normal distribution's mean bias was also significantly greater than that found for the skewed distribution. Finally, $\hat{\Phi}$ was more biased than $\hat{\Phi}(\lambda)$, although the differences for the J and bimodal distributions were negligible.

Moving the location of the cut-off score was expected to have no influence on the coefficients' accuracy. The mean bias associated with each cut-off score can be found in Table 11. These means were based on the results for 10, 15, and 20-item tests. Five-item tests were not included because only one cut-off score was examined for this test length, i.e., the design of the study was not completely crossed. Although the cut-off scores associated with the 15-item tests were not exactly equal to those of the other two test lengths, the researcher felt the slight deviations would not significantly affect the results.

Table 11.--Mean Bias Across Cells of Each Coefficient for Each Cut-off Score.

	Cut-Off Score		
Coefficient	70%	80%	90%
Livingston's $\hat{K}^2(X,T_X)$	.016	-.004	-.003
Brennan & Kane's $\hat{\Phi}(\lambda)$[a]	.038	.032	.035
Subkoviak's $\hat{p}_o$	-.012	-.029	-.012
Huynh's $\hat{p}_o$	-.010	-.014	.027
Subkoviak's $\hat{K}$	-.071	-.069	-.030
Huynh's $\hat{K}$	-.052	-.016	.075

[a]Means for these coefficients were based only on cells within the randomly parallel condition.

As can be seen, the expectation was confirmed for $\hat{K}^2(X,T_X)$, $\hat{\Phi}(\lambda)$, and Subkoviak's $\hat{p}_o$ since changes in the cut-off score did not significantly alter these coefficients' accuracy. However, the biases of Huynh's estimates and Subkoviak's $\hat{K}$ for the 90% cut-off were significantly different from those found for the other two cut-offs. Specifically, Huynh's $\hat{K}$ significantly overestimated kappa for the 90% cut-off, but significantly and moderately underestimated this parameter for cut-offs of 70% and 80%, respectively. Huynh's $\hat{p}_o$ followed a similar pattern, although the bias associated with the 70% cut score was not significant. Subkoviak's $\hat{K}$ significantly underestimated kappa, regardless of cut-off score, but did so significantly less for the 90% cut score. Finally, for Huynh's $\hat{K}$, setting the cut-off score at 70% led to significantly more underestimation than did the 80% cut-off.

Since the main effects hypotheses concerning bias were generally unsupported, a three-way interaction effect among the relevant variables

(i.e., type of parallelism, distribution, and cut-off score) was examined. The bias of each 5-item test was again excluded because only one cut-off score was examined for this test length. The results of this analysis can be seen in Table 12 and are discussed below for each coefficient separately.

Livingston's $\hat{K}^2(X,T_x)$. For the J and bimodal distributions, neither violating the classic parallelism assumption nor moving the cut-off score significantly altered this coefficient's accuracy. On the other hand, the absolute mean biases belonging to the skewed and normal distributions were significantly greater in the randomly parallel condition than in the classically parallel case for cut-off scores located nearest to the distributions' population means. Accounting for both magnitude and direction, altering parallelism conditions significantly changed the bias of $\hat{K}^2(X,T_x)$ for every cut-off score within the skewed distribution and for the 70% cut-off within the normal distribution. In the former case, the differences increased as the cut-off approached the population mean since the mean bias became more negative in the classically parallel case and more positive in the randomly parallel condition. As a matter-of-fact, varying the cut-off score significantly altered the bias in the randomly parallel condition. Significant differences as a function of cut-off score were also evident in the classically parallel condition when the results for the 70% and 90% cut-offs were compared. For the normal distribution, altering the cut-off did not appreciably affect the mean bias in the classically parallel condition. However, given random parallelism, the mean bias associated with the 70% cut-off was very

Table 12.--Mean Bias Across Cells of Each Coefficient for Every Parallelism/Distribution/Cut-off Score Combination.

Coefficient	Type of Parallelism	70% Skewed	J	Bi-modal	Normal
Livingston's $\hat{K}^2(X,T_X)$	Random	.018	.004	-.016	.171
	Classic	-.017	-.008	-.026	.008
Brennan & Kane's $\hat{\Phi}(\lambda)$	Random	.027	.008	.013	.107
Subkoviak's $\hat{p}_o$	Random	-.010	-.010	-.028	.059
	Classic	-.019	-.015	-.026	-.044
Huynh's $\hat{p}_o$	Random	-.017	-.006	-.037	.090
	Classic	-.030	-.015	-.046	-.014
Subkoviak's $\hat{K}$	Random	.000	-.028	-.071	-.037
	Classic	-.205	-.040	-.084	-.106
Huynh's $\hat{K}$	Random	.016	-.012	-.078	.037
	Classic	-.175	-.034	-.124	-.037

Table 12 (cont'd.)

Coefficient	Type of Parallelism	80% Skewed	J	Bi-modal	Normal
Livingston's $\hat{K}^2(\underline{X},\underline{T}_{\underline{X}})$	Random	.050	.002	-.017	-.003
	Classic	-.028	-.008	-.031	.000
Brennan & Kane's $\hat{\Phi}(\lambda)$	Random	.054	.007	.015	.053
Subkoviak's $\hat{p}_{\underline{O}}$	Random	-.009	-.008	-.026	-.032
	Classic	-.035	-.027	-.037	-.062
Huynh's $\hat{p}_{\underline{O}}$	Random	-.009	.016	-.012	.002
	Classic	-.038	-.003	-.032	-.036
Subkoviak's $\hat{\underline{K}}$	Random	.026	-.022	-.048	-.025
	Classic	-.162	-.066	-.089	-.170
Huynh's $\hat{\underline{K}}$	Random	.062	.032	-.007	.056
	Classic	-.109	-.008	-.065	-.093

Table 12 (cont'd.)

Coefficient	Type of Parallelism	90% Skewed	J	Bi-modal	Normal
Livingston's $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$	Random	.088	.002	-.014	-.020
	Classic	-.042	-.007	-.030	-.002
Brennan & Kane's $\hat{\Phi}(\lambda)$	Random	.083	.005	.014	.038
Subkoviak's $\hat{\underline{p}}_{\underline{o}}$	Random	.006	.018	-.003	-.058
	Classic	-.036	-.008	-.005	-.013
Huynh's $\hat{\underline{p}}_{\underline{o}}$	Random	.045	.078	.052	-.065
	Classic	.007	.061	.048	-.021
Subkoviak's $\hat{\underline{K}}$	Random	.061	.027	-.005	-.056
	Classic	-.057	-.027	-.008	-.190
Huynh's $\hat{\underline{K}}$	Random	.150	.151	.108	.024
	Classic	.050	.118	.101	-.114

large and significantly different from that found for the other two cut-offs. Specifically, $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ greatly overestimated its parameter when the cut-off equalled 70%, but fairly accurately estimated $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ for the 80% cut-off, and moderately underestimated $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ given the 90% cut-off.

Although no hypothesis was made concerning the influence of distributional shape, this variable did have an impact. In the randomly parallel condition, the effects varied across cut-off score due to the changes induced by this variable within the normal and skewed distributions. The J distribution resulted in the least bias. As a matter-of-fact, its bias was close to zero, regardless of cut-off score. Although not significant, $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ consistently underestimated its parameter for the bimodal distribution. The relationship between the J and bimodal distributions' results remained fairly consistent across cut-off score. The greatest degree of bias was associated with the normal distribution for the 70% cut-off and with the skewed distribution given the other two cut-offs. In these instances, the bias differed significantly from that corresponding to the other distributions with each one significantly overestimating its population value. The only other significant difference was found between the bimodal and skewed distributions for a 70% cut-off score. $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ underestimated $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ in the former case and overestimated $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ in the latter case.

For the classic parallelism condition, the pattern of mean bias created by changing the distribution was similar across cut-off score. The mean biases for the normal and J distributions were almost zero in every case with the former distribution resulting in no bias for the 80%

cut-off. The mean biases associated with the skewed and bimodal distributions were quite similar. In both cases, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ was consistently underestimated with significant bias occurring for the 80% and 90% cut-offs. The bimodal distribution led to significant underestimation for the 70% cut-off, as well. The biases produced by these distributions were significantly different from those found for the normal distribution, regardless of cut-off score. For the 90% cut-off, the skewed distribution also displayed significantly more bias than the J distribution did. Once again, the relationship between the mean biases of the J and bimodal distributions was fairly consistent across cut-off score.

Brennan and Kane's $\hat{\Phi}(\lambda)$. Since $\hat{\Phi}(\lambda)$ was not computed for classically parallel tests in this study, Table 12 contains the mean bias of this coefficient for every distribution/cut-off score combination. The results almost paralleled those found for $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ in the randomly parallel condition. As a matter-of-fact, in terms of absolute value, the pattern of results for $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\hat{\Phi}(\lambda)$ were, with one exception, nearly identical. In many cases, the actual degrees of bias were very similar. Note, however, that $\hat{\Phi}(\lambda)$, on the average, consistently overestimated its parametric value, while $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ did not. The following discussion elaborates upon the similarities between these coefficients.

Altering the cut-off score within the J and bimodal distributions hardly changed this coefficient's accuracy. However, as the cut-off approached the population means of the other distributions, the mean biases increased. For the skewed distribution, each increase was significant. When the distribution was normal, the mean bias associated with

the 70% cut-off score was significantly greater than that found for the other two cut-offs.

Changes in the frequency distribution also altered the results. The mean biases corresponding to the J distribution were consistently close to zero. The bimodal distribution created slightly more inaccuracy. Because neither of these distributions was affected by cut-off score, the relationship between them remained fairly constant across cut-off score. When the cut-off was 70%, the normal distribution's mean bias was extremely large and significantly different from that found for the other distributions. Contrary to the pattern established by $\hat{K}^2(\underline{X},\underline{T}_{\underline{X}})$, the biases of the skewed and normal distributions were, on the average, comparable, significant, and significantly different from that found for the other two distributions when the cut-off equalled 80%. Finally, given a 90% cut-off, the skewed distribution produced a large mean bias which was significantly greater than the biases of the other distributions. In this situation, the mean bias associated with the normal distribution was also significant as well as significantly greater than the J distribution's mean bias.

Subkoviak's $\hat{p}_{\underline{Q}}$. In terms of absolute value, violating the classic parallelism assumption did not significantly affect the accuracy of Subkoviak's $\hat{p}_{\underline{Q}}$ for the J and bimodal distributions. If one considers direction, however, type of parallelism did significantly alter the J distribution's results when the cut-off was 90%; on the average, the random parallelism situation led to overestimation, while its classically parallel counterpart produced a fairly accurate estimate.

Contrary to the hypothesis, the skewed distribution's absolute mean bias was greater when the classic parallelism assumption was valid. As the cut-off approached this distribution's population mode (mean), the differences in the mean bias of the classically and randomly parallel conditions increased since the bias became more negative in the former case and less negative in the latter case. In the latter condition, the bias was even slightly positive for the 90% cut-off. However, whether or not direction was taken into account, violating the classic parallelism assumption significantly altered only the biases corresponding to the 80% and 90% cut-offs.

Although neither type of parallelism was consistently associated with less bias, significant differences also occurred for the normal distribution. For the 80% cut-off, the absolute mean bias was greater in the classically parallel situation, while the opposite was true for the 90% cut-off. Taking direction into account, violating the classic parallelism assumption significantly altered the results for all cut-off scores within this distribution. For the 80% and 90% cut-off scores, $\hat{\underline{p}}_{\underline{o}}$ consistently underestimated its parametric value. The bias corresponding to the 70% cut-off was negative in the classically parallel condition but positive in the randomly parallel situation.

Keeping type of parallelism constant, changing the cut-off score did not significantly alter the mean biases within the beta-binomial distributions, except in the case of randomly parallel J distributed tests. In this instance, the mean bias for the 90% cut-off was positive, while the mean biases for the 70% and 80% cut-offs were slightly negative. When the distribution was either bimodal or normal, moving the cut-off did

lead to significant differences. Specifically, given classic parallelism, the 80% cut-off produced much more underestimation than did the 90% cut-off. In addition, for the normal distribution, the 70% cut-off resulted in significantly more negative bias than did the 90% cut-off. The accuracy of the reliability estimates for randomly parallel tests was also significantly affected. When the distribution was bimodal, Subkoviak's $\hat{p}_o$ largely underestimated p_o for a 70% cut-off but fairly accurately estimated the population value given the 90% cut-off. For the normal distribution, Subkoviak's coefficient largely overestimated p_o when the cut-off was 70% while largely underestimating p_o for higher cut-off scores. Also, the 90% cut-off produced significantly more underestimation than did the 80% cut-off.

For the 70% and 80% cut-off scores, the patterns of results formed by changing the distributional shape were similar and partially supported the hypothesis that the normal and bimodal distributions would produce more bias than the beta-binomial distributions. Given random parallelism, the biases corresponding to the J and skewed distributions were low, negative, and approximately equal. The bimodal distribution produced significant underestimation, but the results were not significantly different from those found for the J and skewed distributions. Although differing in direction, the normal distribution was significantly biased for both cut-off scores and, for the cut-off closest to its population mode (mean), significantly more biased than the other distributions. When the cut-off equalled 80%, the bias found for the normal distribution was also much worse than that found for the beta-binomial distributions, but the differences did not attain significance. Given classic parallelism, the J distribution once again produced the

least bias with significant underestimation occurring for the 80% cut-off. The skewed and bimodal distributions resulted in slightly more negative bias which, therefore, also reached significance for the 80% cut-off. For the latter distribution, the extent of underestimation was also greater than -.025 when the cut-off was 70%. When the distribution was normal, significant underestimation occurred for both cut-off scores. In support of the hypothesis, the differences between the normal distribution's results and those of the beta-binomial distributions attained significance. The mean biases associated with the normal and bimodal distributions also differed significantly for the 80% cut-off.

The pattern of results for the 90% cut-off was somewhat different. Unexpectedly, the bimodal distribution, on the average, produced fairly accurate estimates in both parallelism conditions. For the randomly parallel situation, the negative bias associated with the normal distribution was significant and, in terms of absolute value, significantly greater than that found for the other distributions. However, such was not the case for the classically parallel condition. As a matter-of-fact, the skewed distribution claimed the greatest bias which was significantly negative as well as significantly greater than that found for the J and bimodal distributions.

Huynh's $\hat{p}_o$. Similar to the results found for Subkoviak's $\hat{p}_o$, violating the classic parallelism assumption did not significantly affect the bias within the J and bimodal distributions. When the distribution was skewed, type of parallelism did significantly alter the results for the 80% and 90% cut-offs. These differences were significant whether or not direction was taken into account. The direction and degrees of bias

corresponding to the 80% cut-off were almost exactly the same as those found for Subkoviak's $\hat{p}_o$ with the classic parallelism situation resulting in more underestimation. However, for the 90% cut-off, Huynh's $\hat{p}_o$ was significantly more biased in the randomly parallel condition, while Subkoviak's estimate produced more bias in the classically parallel condition. In fact, Huynh's coefficient significantly overestimated the parameter in the randomly parallel condition but provided a fairly accurate estimate in the classically parallel condition. Finally, whether or not one considers direction, the biases associated with each parallelism condition within the normal distribution were significantly different from each other, regardless of cut-off score. Once again, both p_o estimates followed a similar pattern. For cut-offs of 70% and 90%, the absolute mean bias associated with random parallelism was greater than that found in the classically parallel condition, while the opposite held true for the 80% cut-off. The mean bias was consistently negative in the classic parallelism condition. However, in the random parallelism situation, the mean bias was highly positive, virtually zero, and highly negative for the 70%, 80%, and 90% cut-offs, respectively.

Changes in the cut-off score significantly impacted the mean bias within every distribution. For those distributions with their population mode (mean) close to 90%, the mean bias corresponding to this extreme cut-off differed significantly from that found for the other cut-off scores. Generally, the mean biases for the 90% cut-off were significantly positive, while the mean biases associated with the other cut-off scores ranged from slightly positive to significantly negative. One exception to this trend occurred when estimating the reliability of

classically parallel tests having skewed distributions. In this case, the mean bias for the 90% cut-off was close to zero. Among the three distributions, the bimodal produced the only significant difference between the 70% and 80% cut-offs; given random parallelism, significantly more negative bias was found for the former than for the 80% cut-off.

On the other hand, within the normal distribution, altering the cut-off had no major effect in the classic parallelism condition. In the case of random parallelism, the results for the various cut-off scores were all significantly different from each other. As noted previously, the 70% cut-off led to significant overestimation as it had done for most of the other coefficients, while the 80% and 90% cut-off scores resulted in a fairly accurate estimate and a significant negative bias, respectively.

The hypothesis that the normal and bimodal distributions would produce more bias than the beta-binomial distributions was generally unsupported. Although the type of distribution significantly affected the direction and/or extent of bias, no consistent pattern could be found either across cut-off score or parallelism condition.

Subkoviak's $\hat{K}$. Contrary to previous results, type of parallelism significantly affected the accuracy of Subkoviak's $\hat{K}$ within every distribution. In terms of absolute value, the J and bimodal distributions were sensitive to this variable when the cut-off was 80%; $\hat{K}$ was more negatively biased in the classically parallel condition. When direction was considered, parallelism produced an additional significant effect for the J distribution. Specifically, when the cut-off was 90%, the biases associated with the two types of parallelism were equal but opposite in

direction. As usual, parallelism significantly affected the results associated with the skewed and normal distributions. In terms of absolute value, the differences were significant for all cut-off scores within the normal distribution and for the 70% and 80% cut-off scores within the skewed distribution. In all these cases, the classic parallelism condition produced more bias than did its randomly parallel counterpart. When direction was considered, all respective comparisons within these two distributions were significant. For classically parallel tests having skewed distributions, $\hat{\underline{K}}$ consistently underestimated its parameter. However, for randomly parallel skewed tests, $\hat{\underline{K}}$ was, on the average, unbiased when the cut-off was 70% and positively biased given the other cut-offs. When the tests were normally distributed, $\hat{\underline{K}}$ underestimated the population value, regardless of cut-off score and type of parallelism.

Cut-off score also had a pervasive effect. For the three distributions having their population modes (means) near 90%, the mean biases associated with this extreme cut-off were, in general, significantly less negative than that found for the other two cut-offs. Two very distinct deviations from this trend occurred for the skewed and J distributions within the randomly parallel condition. For the former distribution, the mean bias was either zero or positive and increased significantly as the cut-off approached 90%. For the J distribution, moving the cut-off affected the bias' direction but not its magnitude in that $\hat{\underline{K}}$ overestimated kappa for the 90% cut-off and almost equally underestimated this parameter for the other cut-offs. In addition, given classic parallelism, the 80% cut-off produced significantly greater

underestimation than did the 70% cut-off for the J distribution, while the opposite occurred for the skewed distribution.

The pattern of results formed by moving the cut-off was quite different for the normal distribution. In the randomly and classically parallel situations, the 90% cut-off produced more underestimation than did the 80% and 70% cut-offs, respectively. In addition, given classic parallelism, the 80% cut-off resulted in significantly more negative bias than did the 70% cut-off.

Generally, the hypothesis that $\hat{K}$ would be more accurate for beta-binomial distributions was not supported. Once again, the pattern of results formed by altering the distribution varied across cut-off score. However, in the classically parallel condition, $\hat{K}$ significantly underestimated kappa for every distribution and cut-off score, except one; for the bimodal distribution, the bias found when the cut-off was 90% was close to zero. In the randomly parallel condition, the degree of bias was generally significant but varied in direction. However, for the skewed and bimodal distributions, the mean bias was zero or close to zero for the 70% and 90% cut-off scores, respectively.

Huynh's $\hat{K}$. With relatively few exceptions, the bias of Huynh's $\hat{K}$ was significantly affected when any of the variables in the study changed values. Moreover, the results did not follow any pattern, making interpretation very difficult. Therefore, the following observations are not as specific as those presented for the other coefficients.

Across all distributions and cut-off scores, the absolute mean biases within the parallelism conditions were comparable in only four cases. For the 90% cut-off within the J distribution, $\hat{K}$ produced significantly more overestimation in the randomly parallel condition. On

the other hand, when the distribution was bimodal, the bias was significantly more negative in the classically, as opposed to the randomly parallel, condition when the cut-off was either 70% or 80%. In terms of absolute value, the skewed distribution also produced significantly more bias for the classically parallel situation given cut-off scores of 70% and 80%. However, when the cut-off equalled 90%, the randomly parallel condition produced significantly more bias. Finally, the absolute mean bias for classically parallel normally distributed tests was greater than that found for their randomly parallel counterparts for cut-off scores of 80% and 90%.

Violating the classic parallelism assumption also affected the direction as well as the magnitude of the bias, especially for the skewed and normal distributions. For both these distributions, $\underline{\hat{K}}$ was positive in the random parallelism condition and, generally, negative in the classic parallelism condition.

For the three distributions having their modes (means) near 90%, the bias associated with this extreme cut-off differed significantly from that found for the other cut-offs. Specifically, the mean biases for the 90% cut-off were significantly positive, while the mean biases corresponding to the other cut-offs ranged from significantly positive to significantly negative. In all three distributions, the results for the 70% and 80% cut-offs also differed significantly, regardless of parallelism. As the cut-off moved from 70% to 90%, the bias moved toward overestimation.

Changing the cut-off did not affect the normal distribution as drastically. In the randomly parallel condition, the bias for the 80% cut-off was significantly more positive than that found for the 90% cut-

off. When the tests were classically parallel, the 70% cut-off produced significantly less negative bias than did the 80% and 90% cut-off scores.

The hypothesis that $\hat{\underline{K}}$ would be more accurate for the skewed and J distributions than for the other two distributions was generally not supported. Once again, the pattern of results was inconsistent across cut-off score. In general, the degrees of bias were significant. However, for the 80% cut-off, the biases associated with the J and bimodal distributions were close to zero for the classically and randomly parallel conditions, respectively.

Sampling Variability

The variability of each coefficient across samples was predicted to be inversely related to the test length and the sample size. For each coefficient, Tables 13 and 14 present the weighted mean standard deviation across cells associated with each test length and sample size, respectively. Clearly, the results support the hypothesis, i.e., the sampling variability decreased as the test length increased as well as when the sample size increased. As can be seen, Subkoviak's coefficients were more variable than Huynh's coefficients for every test length and cut-off score. $\hat{\Phi}$ appeared to be more unstable than $\hat{\Phi}(\lambda)$.

Table 15 contains the mean standard deviation for every test length/sample size combination. No significant interaction effects were evident.

Table 13.--Mean Standard Deviation Across Cells of Each Coefficient for Each Test Length.

	Test Length		
Coefficient	10	15	20
Livingston's $\hat{K}^2(X,T_X)$	.049	.039	.032
Brennan & Kane's $\hat{\Phi}(\lambda)$[a]	.042	.028	.019
Brennan & Kane's $\hat{\Phi}$[a]	.050	.037	.026
Subkoviak's $\hat{p}_o$	.035	.033	.030
Huynh's $\hat{p}_o$	.029	.025	.022
Subkoviak's $\hat{K}$	.089	.085	.078
Huynh's $\hat{K}$	.073	.064	.058

[a]The mean standard deviations for these coefficients were based only on cells within the randomly parallel condition.

Table 14.--Mean Standard Deviation Across Cells of Each Coefficient for Each Sample Size.

	Sample Size		
Coefficient	25	35	50
Livingston's $\hat{K}^2(X,T_X)$	.050	.043	.042
Brennan & Kane's $\hat{\Phi}(\lambda)$[a]	.040	.036	.031
Brennan & Kane's $\hat{\Phi}$[a]	.050	.043	.037
Subkoviak's $\hat{p}_o$	.038	.033	.030
Huynh's $\hat{p}_o$	.030	.025	.024
Subkoviak's $\hat{K}$	.099	.085	.077
Huynh's $\hat{K}$	.078	.067	.063

[a]The mean standard deviations for these coefficients were based only on cells within the randomly parallel condition.

Table 15.--Mean Standard Deviation Across Cells for Each Sample Size/Test Length Combination.

Coefficient	Test Length	Sample Size 25	35	50
Livingston's $\hat{K}^2(X,T_X)$	10	.055	.048	.046
	15	.043	.037	.038
	20	.036	.031	.029
Brennan & Kane's $\hat{\Phi}(\lambda)$[a]	10	.047	.044	.037
	15	.031	.027	.026
	20	.022	.019	.016
Brennan & Kane's $\hat{\Phi}$[a]	10	.057	.051	.042
	15	.043	.035	.033
	20	.031	.025	.022
Subkoviak's $\hat{p}_o$	10	.039	.034	.032
	15	.038	.031	.029
	20	.034	.030	.027
Huynh's $\hat{p}_o$	10	.032	.028	.026
	15	.028	.023	.022
	20	.025	.021	.019
Subkoviak's $\hat{K}$	10	.100	.088	.080
	15	.097	.081	.076
	20	.091	.077	.068
Huynh's $\hat{K}$	10	.081	.073	.067
	15	.073	.061	.060
	20	.067	.057	.052

[a]The mean standard deviations for these coefficients were based only on cells within the randomly parallel condition.

DISCUSSION

Although all the coefficients in this study, except for $\hat{\Phi}(\lambda)$, were derived under the assumption of classic parallelism, they were in many cases robust to violation of this assumption, i.e., type of parallelism did not significantly alter the absolute mean bias. Specifically, for $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ and the $\underline{p}_{\underline{Q}}$ estimates, this variable had no significant effect when the distributions were either J-shaped or bimodal. However, type of parallelism, in general, affected the absolute mean bias when the distributions were either skewed or normal. These findings can perhaps be explained by examining the item characteristics which must be present to form each distribution. If the domain score distribution is either J-shaped or bimodal, the domain must consist of items having fairly homogeneous $\underline{p}$ values and high item intercorrelations. On the other hand, items within a domain having either a skewed or a normal distribution are more heterogeneous and have lower item intercorrelations. When items are randomly chosen to construct alternate forms, some or all of the statistics computed from one test are more likely to adequately represent the characteristics of the other form when the item domain is more homogeneous. In other words, the relationship between randomly parallel tests derived from a homogeneous, in contrast to a heterogeneous, item domain more closely resembles that found between classically parallel tests. In addition, for $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$, coefficient alpha is probably a better estimate of the alternate form reliability when the domain is homogeneous, leading to more accurate estimation of

$\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ for randomly parallel tests having a J or bimodal distribution. Given these facts, one would expect type of parallelism to have a greater effect when alternate forms are derived from a heterogeneous item domain, i.e., from an item domain having either a skewed or a normal distribution.

Type of parallelism did affect the absolute mean bias of the kappa estimates, regardless of distribution. However, the effects were somewhat less pervasive for the J and bimodal distributions.

When parallelism did significantly alter the absolute mean biases, the random parallelism condition did not always result in the most bias. For example, except in one case, Subkoviak's coefficients displayed greater bias in the classic parallelism situation. This latter result can perhaps be understood by examining Subkoviak's formula more closely. Specifically, notice that using a regression estimate of true score causes regression toward the mean which becomes more severe as $\underline{KR}$-20 decreases. When the distribution was either normal or skewed in this study, $\underline{KR}$-20 was likely to be fairly low. The resultant regression toward the mean may have caused the alternate form population value within the classically parallel condition to be severely underestimated, leading to a greater degree of bias for the classically parallel situation. As predicted, for $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$, the random parallelism condition did produce the most bias. Finally, the type of parallelism associated with more bias varied for Huynh's estimates.

Cut-off score had a significant effect on all the coefficients. For $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\hat{\Phi}(\lambda)$, the effects were found predominantly within the random parallelism condition when the distributions were either

skewed or normal. For the most part, these biases tended to increase as the cut-off approached the mean; the biases were positive when the cut-off was located near the population mean. Note that as the cut-off moves close to the mean, the difference between the mean and the cut-off has less of an impact on the value of these coefficients. When the cut-off almost equals the mean, the squared-error agreement coefficients are approximately equal to KR-20 or KR-21. Therefore, the present results indicate that as the difference between the mean and the cut-off becomes less influential and the norm-referenced reliability coefficient accounts more for the magnitude of $\hat{K}^2(X,T_X)$ and $\hat{\Phi}(\lambda)$, the bias becomes significantly more positive for randomly parallel tests having a skewed or a normal distribution. For the J and bimodal distributions, the bias of $\hat{K}^2(X,T_X)$ and $\hat{\Phi}(\lambda)$ did not significantly change as a function of cut-off score. Because of these distributions' homogeneous item domains, KR-20 is probably a fairly accurate estimate of the population coefficient used in these formulas.

No general statements can be made about the effect of cut-off score on the threshold agreement coefficients since the results varied widely. The most consistent finding occurred for Huynh's coefficients. Specifically, for distributions having their population mode (mean) near 90%, the bias associated with this extreme cut-off score was significantly positive. Significant overestimation may occur in this case because, according to the binomal error model, the standard deviation around an extreme true score is smaller than that around non-extreme scores. However, the data in this study do not conform

exactly to the binomial error model and, therefore, scores may be more variable than what is predicted by this model. Since scores within these distributions cluster about the 90% cut-off, this increased variability will have more of an effect in decreasing the population values, leading to overestimation by Huynh's coefficients.

The hypothesis that the p_o and kappa estimates would be less biased for the beta-binomial distributions than for the bimodal and normal distributions was, generally, not supported. A possible explanation for this finding is that the J and skewed distributions in this study did not conform closely enough to members of the beta-binomial family. In addition, the normal and skewed distributions may not have been different enough since the normal distribution was actually somewhat skewed.

SUMMARY AND CONCLUSIONS

Due to the variable results found in this study, no general rules can be offered for choosing between coefficients falling within each category (e.g., uncorrected threshold agreement coefficients) of Figure 2. For each parallelism/distribution/cut-off score combination, Table 16 indicates the direction of bias produced by each coefficient. If a coefficient had a mean bias less than .025 in either direction, the coefficient was considered to be unbiased. Recommendations about which coefficient to use in each of these cells can be made and are presented in Table 17. Two criteria were used in choosing a coefficient in each case:

(1) when the biases were in the same direction, the coefficient with the least bias was selected;

(2) when the biases were opposite in direction, the negatively biased coefficient was chosen, unless the positively biased coefficient was fairly accurate (i.e., had a bias near zero) or much more accurate than its competitor.

The latter situation occurred only once where Huynh's $\hat{\underline{K}}$ had a positive, but nonsignificant, bias, while Subkoviak's $\hat{\underline{K}}$ had a significant negative bias.

Several other points about Table 17 should be mentioned. First, sampling variability was not taken into account in making these recommendations. There were two reasons for taking this course of action: (1) the bias of an estimator is more important than its

Table 16.--Direction of Bias of Each Coefficient for Each Parallelism/Distribution/Cut-off Score Combination

Coefficient	Type of Parallelism	70% Skewed	J	Bi-modal	Normal
Livingston's $\hat{K}^2(X,T_x)$	Random	No Bias	No Bias	No Bias	Over-est.
	Classic	No Bias	No Bias	Under-est.	No Bias
Brennan & Kane's $\hat{\Phi}(\lambda)$	Random	Over-est.	No Bias	No Bias	Over-est.
Brennan & Kane's $\hat{\Phi}$	Random	Over-est.	No Bias	No Bias	Over-est.
Subkoviak's $\hat{p}_o$	Random	No Bias	No Bias	Under-est.	Over-est.
	Classic	No Bias	No Bias	Under-est.	Under-est.
Huynh's $\hat{p}_o$	Random	No Bias	No Bias	Under-est.	Over-est.
	Classic	Under-est.	No Bias	Under-est.	No Bias
Subkoviak's $\hat{K}$	Random	No Bias	Under-est.	Under-est.	Under-est.
	Classic	Under-est.	Under-est.	Under-est.	Under-est.
Huynh's $\hat{K}$	Random	No Bias	No Bias	Under-est.	Over-est.
	Classic	Under-est.	Under-est.	Under-est.	Under-est.

Table 17.--Recommended Corrected/Uncorrected Squared Error and Threshold Agreement Coefficients for Each Parallelism/Distribution/Cut-off Score Combination.

		Random Parallelism				Classic Parallelism				
		Type of Coefficient					Type of Coefficient			
		Uncorrected Squared Error	Corrected Squared Error	Uncorrected Threshold (p_Q)[a]	Corrected Threshold (K)		Uncorrected Squared Error	Corrected Squared Error	Uncorrected Threshold (p_Q)[a]	Corrected Threshold (K)
Skewed	70	$K^2(X,T_X)$	Φ	Subkoviak's	Subkoviak's	Skewed	70	$K^2(X,T_X)$	Φ	Subkoviak's	Huynh's
	80	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's/ Huynh's	Subkoviak's		80	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Huynh's
	90	$\Phi(\lambda)$	Φ	Subkoviak's	Subkoviak's		90	$K^2(X,T_{\bar{X}})$	Φ	Huynh's	Subkoviak's
J-Shaped	70	$K^2(X,T_X)$	Φ	Huynh's	Huynh's	J-Shaped	70	$K^2(X,T_X)$	Φ	Subkoviak's/ Huynh's	Huynh's
	80	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Subkoviak's		80	$K^2(X,T_{\bar{X}})$	Φ	Huynh's	Huynh's
	90	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Subkoviak's		90	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Subkoviak's
Bimodal	70	$K^2(X,T_X)$	Φ	Subkoviak's	Subkoviak's	Bimodal	70	$K^2(X,T_X)$	Φ	Subkoviak's	Subkoviak's
	80	$K^2(X,T_{\bar{X}})$	Φ	Huynh's	Huynh's		80	$K^2(X,T_{\bar{X}})$	Φ	Huynh's	Huynh's
	90	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Subkoviak's		90	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Subkoviak's
Normal	70	$\Phi(\lambda)$	Φ	Subkoviak's	Subkoviak's	Normal	70	$K^2(X,T_X)$	Φ	Huynh's	Huynh's
	80	$K^2(X,T_X)$	Φ	Huynh's	Subkoviak's		90	$K^2(X,T_{\bar{X}})$	Φ	Huynh's	Huynh's
	90	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Huynh's		90	$K^2(X,T_{\bar{X}})$	Φ	Subkoviak's	Huynh's

[a]Marshall's estimate, as computed in this study, is equal to Subkoviak's estimate and can, therefore, be used wherever the latter is recommended.

sampling variability in determining the estimator's adequacy; and (2) Tables 13, 14, and 15 indicate that the differences in stability of each coefficient within a particular category are not very large. Second, there were two instances where Subkoviak's and Huynh's coefficients were equally biased, and, consequently, both were listed. However, since Huynh's coefficients appeared to be slightly more stable, one might want to select his estimates. Third, even though $\hat{\Phi}$ significantly overestimated its parametric value when the distribution was either skewed or normal, this coefficient is the only available corrected squared-error formula and was, therefore, recommended in every situation. Finally, although either $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ or $\hat{\Phi}(\lambda)$ was recommended in each case, one must remember that these two coefficients are not really comparable because they do not estimate the same population value. Specifically, $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ measures the reliability associated with a particular test, while $\hat{\Phi}(\lambda)$ indicates the reliability of any set of items randomly selected from a domain. Because of the latter fact, all tests which can possibly be constructed from an item domain must be classically parallel for the classic parallelism assumption to be valid. This situation is unlikely to occur, unless all items within the domain have equal $\underline{p}$ values. When this situation does occur, $\underline{K}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\Phi(\lambda)$ are equal, and one can directly compare the accuracy of $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ and $\hat{\Phi}(\lambda)$. However, since this study did not contain a domain of items having equal $\underline{p}$ values, $\hat{\Phi}(\lambda)$ could not be computed in the classic parallelism condition. Therefore, $\hat{\underline{K}}^2(\underline{X},\underline{T}_{\underline{X}})$ was consistently recommended as the

appropriate uncorrected squared-error coefficient within the classic parallelism condition.

Finally, mention should also be made of two other methods of estimating reliability; Subkoviak and Wilcox (1978) and Livingston and Wingersky (1979) introduced mastery coefficients which measure the extent of agreement between the observed score and the estimated true score. The former coefficient uses a threshold loss function. Livingston and Wingersky's index reflects the size of the misclassification error but does not use a squared error loss function. Since reliability is really concerned with accurately estimating an examinee's true score or true classification, these indices deserve considerable attention.

APPENDICES

APPENDIX A1

Mean Bias and Standard Deviation of Livingston's $\hat{K}^2(X,T_X)$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
5	4	-.015 / .101	-.017 / .039	.022 / .101	.04 / .157	-.083 / .14	-.022 / .05	-.092 / .091	.103 / .128
10	7	.023 / .043	.004 / .021	-.021 / .03	.226 / .119	-.016 / .042	-.008 / .016	-.032 / .024	.027 / .115
	8	.064 / .078	.002 / .022	-.021 / .039	-.023 / .099	-.017 / .079	-.009 / .017	-.038 / .029	-.006 / .122
	9	.116 / .119	.002 / .019	-.018 / .036	-.041 / .052	.003 / .126	-.009 / .016	-.04 / .034	-.012 / .062
15	11	.015 / .035	.002 / .013	-.01 / .016	.151 / .129	-.023 / .033	-.005 / .012	-.016 / .015	.019 / .114
	12	.028 / .063	.001 / .014	-.01 / .019	-.011 / .096	-.031 / .057	-.006 / .013	-.017 / .017	.007 / .092
	14	.041 / .094	.001 / .012	-.008 / .017	-.026 / .028	-.031 / .113	-.005 / .011	-.017 / .018	.003 / .026

APPENDIX A1 (cont'd.)

Mean Bias and Standard Deviation of Livingston's $\hat{K}^2(X,T_X)$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.02 / .019	.005 / .01	-.001 / .01	.127 / .135	-.014 / .019	-.007 / .01	-.01 / .01	-.016 / .126
	16	.052 / .044	.004 / .012	.000 / .014	-.014 / .082	-.021 / .04	-.007 / .011	-.012 / .014	-.029 / .071
	18	.078 / .073	.002 / .011	-.002 / .014	-.013 / .026	-.02 / .075	-.006 / .01	-.012 / .016	-.012 / .024

APPENDIX A2

Mean Bias and Standard Deviation of Livingston's $\hat{K}^2(X,T_X)$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.048 / .114	-.018 / .031	.018 / .092	.025 / .115	-.089 / .079	-.022 / .029	-.117 / .087	.063 / .101
10	7	.022 / .032	.006 / .016	-.032 / .036	.235 / .097	-.012 / .024	-.009 / .009	-.043 / .027	.008 / .095
	8	.073 / .059	.006 / .015	-.033 / .058	.008 / .091	-.019 / .042	-.009 / .01	-.05 / .029	.007 / .096
	9	.143 / .096	.006 / .012	-.027 / .057	-.022 / .046	-.044 / .123	-.008 / .009	-.048 / .035	.000 / .047
15	11	.016 / .023	.004 / .009	-.019 / .026	.179 / .112	-.017 / .018	-.003 / .006	-.024 / .016	.033 / .083
	12	.034 / .036	.004 / .009	-.019 / .035	.028 / .094	-.026 / .027	-.003 / .006	-.025 / .021	.031 / .078
	14	.056 / .061	.003 / .007	-.012 / .028	-.012 / .027	-.08 / .125	-.003 / .005	-.023 / .024	.012 / .025

APPENDIX A2 (cont'd.)

Mean Bias and Standard Deviation of Livingston's $\hat{K}^2(X,T_X)$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal	Classically Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
20	14	.017 / .014	.006 / .007	-.006 / .015	.114 / .101	-.014 / .015	-.006 / .006	-.016 / .011	-.003 / .086
	16	.05 / .025	.006 / .007	-.006 / .024	.002 / .086	-.027 / .031	-.005 / .006	-.018 / .018	.004 / .054
	18	.081 / .046	.005 / .006	-.005 / .024	-.003 / .03	-.044 / .082	-.004 / .005	-.016 / .021	.002 / .019

APPENDIX A3

Mean Bias and Standard Deviation of Livingston's $\hat{K}^2(X,T_X)$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
5	4	-.044 / .068	-.024 / .036	.008 / .07	.028 / .135	-.118 / .119	-.035 / .04	-.116 / .093	.044 / .071
10	7	.025 / .027	.002 / .019	-.032 / .024	.217 / .107	-.018 / .034	-.013 / .014	-.047 / .024	-.018 / .096
	8	.074 / .039	-.001 / .021	-.036 / .037	-.025 / .095	-.035 / .072	-.014 / .015	-.062 / .037	-.027 / .092
	9	.146 / .044	-.001 / .02	-.031 / .037	-.038 / .043	-.049 / .113	-.013 / .014	-.063 / .05	-.018 / .041
15	11	.012 / .022	.000 / .013	-.019 / .018	.162 / .114	-.023 / .026	-.008 / .009	-.029 / .015	.02 / .099
	12	.029 / .032	-.002 / .014	-.022 / .024	.000 / .109	-.04 / .045	-.009 / .01	-.034 / .024	.015 / .09
	14	.056 / .047	-.002 / .012	-.017 / .021	-.021 / .029	-.07 / .096	-.008 / .01	-.031 / .027	.006 / .022

APPENDIX A3 (cont'd.)

Mean Bias and Standard Deviation of Livingston's $\hat{K}^2(X,T_X)$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.016 / .015	.003 / .009	-.006 / .011	.135 / .089	-.017 / .017	-.009 / .007	-.018 / .009	.009 / .082
	16	.044 / .027	.002 / .011	-.008 / .018	.002 / .07	-.035 / .039	-.009 / .009	-.023 / .018	-.001 / .055
	18	.077 / .037	.001 / .01	-.009 / .018	-.006 / .023	-.046 / .068	-.008 / .009	-.022 / .022	-.002 / .017

APPENDIX A4

Mean Bias and Standard Deviation of Brennan and Kane's $\hat{\Phi}(\lambda)$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	.101 / .102	.019 / .037	-.025 / .093	.015 / .155
10	7	.019 / .036	.004 / .016	.005 / .024	.078 / .114
	8	.041 / .065	.002 / .017	.007 / .031	.034 / .09
	9	.063 / .10	.002 / .015	.009 / .029	.029 / .045
15	11	.035 / .023	.01 / .009	.022 / .01	.072 / .101
	12	.054 / .044	.008 / .01	.024 / .012	.019 / .076
	14	.072 / .066	.006 / .008	.02 / .011	.021 / .022
20	14	.028 / .011	.01 / .005	.024 / .005	.153 / .085
	16	.06 / .025	.009 / .006	.028 / .007	.071 / .053
	18	.09 / .042	.007 / .006	.024 / .007	.049 / .017

APPENDIX A5

Mean Bias and Standard Deviation of Brennan and Kane's $\hat{\Phi}(\lambda)$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	.065 / .121	.019 / .028	-.028 / .092	.006 / .109
10	7	.017 / .028	.006 / .013	-.004 / .032	.088 / .091
	8	.047 / .05	.006 / .012	-.002 / .049	.064 / .08
	9	.084 / .078	.005 / .01	.002 / .047	.045 / .039
15	11	.035 / .016	.011 / .006	.016 / .019	.096 / .083
	12	.058 / .025	.01 / .006	.018 / .025	.051 / .068
	14	.083 / .041	.008 / .005	.017 / .019	.032 / .019
20	14	.026 / .008	.011 / .004	.021 / .009	.147 / .063
	16	.058 / .015	.01 / .004	.025 / .014	.083 / .051
	18	.091 / .026	.009 / .003	.022 / .013	.055 / .017

APPENDIX A6

Mean Bias and Standard Deviation of Brennan and Kane's $\hat{\Phi}(\lambda)$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	.077 / .069	.014 / .032	-.032 / .067	.021 / .126
10	7	.021 / .023	.003 / .015	-.002 / .02	.073 / .098
	8	.052 / .033	.001 / .016	-.003 / .03	.034 / .085
	9	.091 / .036	.000 / .015	.000 / .03	.031 / .037
15	11	.034 / .015	.008 / .008	.017 / .012	.089 / .081
	12	.056 / .022	.006 / .009	.017 / .017	.033 / .078
	14	.084 / .032	.005 / .008	.014 / .015	.026 / .021
20	14	.026 / .008	.009 / .005	.022 / .006	.161 / .056
	16	.056 / .015	.008 / .005	.024 / .01	.082 / .043
	18	.091 / .02	.006 / .005	.021 / .01	.053 / .014

APPENDIX A7

Mean Bias and Standard Deviation of Brennan and Kane's $\hat{\Phi}$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	.101 / .148	.023 / .035	-.023 / .089	.151 / .148
10	7	.069 / .093	.004 / .017	.01 / .03	.121 / .096
	8	.069 / .093	.004 / .017	.01 / .03	.121 / .096
	9	.069 / .093	.004 / .017	.01 / .03	.121 / .096
15	11	.08 / .072	.002 / .009	.024 / .012	.128 / .087
	12	.08 / .072	.002 / .009	.024 / .012	.128 / .087
	14	.08 / .072	.002 / .009	.024 / .012	.128 / .087
20	14	.092 / .04	.01 / .006	.028 / .007	.162 / .076
	16	.092 / .04	.01 / .006	.028 / .007	.162 / .076
	18	.092 / .04	.01 / .006	.028 / .007	.162 / .076

APPENDIX A8

Mean Bias and Standard Deviation of Brennan and Kane's $\hat{\Phi}$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	.065 / .158	.023 / .027	-.029 / .083	.11 / .098
10	7	.086 / .072	.007 / .012	-.001 / .05	.128 / .072
	8	.086 / .072	.007 / .012	-.001 / .05	.128 / .072
	9	.086 / .072	.007 / .012	-.001 / .05	.128 / .072
15	11	.09 / .04	.011 / .006	.018 / .025	.149 / .072
	12	.09 / .04	.011 / .006	.018 / .025	.149 / .072
	14	.09 / .04	.011 / .006	.018 / .025	.149 / .072
20	14	.093 / .024	.011 / .004	.025 / .014	.159 / .059
	16	.093 / .024	.011 / .004	.025 / .014	.159 / .059
	18	.093 / .024	.011 / .004	.025 / .014	.159 / .059

APPENDIX A9

Mean Bias and Standard Deviation of Brennan and Kane's $\hat{\Phi}$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	.069 / .092	.019 / .034	-.034 / .061	.143 / .13
10	7	.093 / .035	.003 / .017	-.002 / .03	.116 / .086
	8	.093 / .035	.003 / .017	-.002 / .03	.116 / .086
	9	.093 / .035	.003 / .017	-.002 / .03	.116 / .086
15	11	.09 / .032	.009 / .009	.017 / .016	.14 / .076
	12	.09 / .032	.009 / .009	.017 / .016	.14 / .076
	14	.09 / .032	.009 / .009	.017 / .016	.14 / .076
20	14	.091 / .02	.01 / .005	.024 / .01	.169 / .054
	16	.091 / .02	.01 / .005	.024 / .01	.169 / .054
	18	.091 / .02	.01 / .005	.024 / .01	.169 / .054

APPENDIX A10

Mean Bias and Standard Deviation of Marshall's $\hat{p}_Q$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.038 / .046	-.04 / .038	-.07 / .054	.09 / .06	-.05 / .052	-.028 / .048	-.101 / .052	-.009 / .05
10	7	-.013 / .035	-.018 / .028	-.035 / .025	.126 / .055	-.026 / .032	-.019 / .022	-.042 / .023	-.023 / .041
	8	-.006 / .046	-.004 / .031	-.028 / .032	-.039 / .037	-.037 / .045	-.032 / .029	-.04 / .029	-.079 / .044
	9	-.001 / .075	.024 / .034	-.024 / .039	-.101 / .051	-.03 / .064	-.007 / .042	.005 / .037	-.04 / .054
15	11	-.003 / .025	-.006 / .025	-.015 / .025	.05 / .059	-.028 / .031	-.015 / .025	-.015 / .022	-.032 / .047
	12	-.008 / .037	-.006 / .028	-.013 / .027	-.016 / .044	-.036 / .039	-.02 / .03	-.028 / .028	-.038 / .048
	14	.004 / .071	.026 / .045	.02 / .045	-.051 / .051	-.017 / .061	.013 / .041	.035 / .032	-.014 / .038

APPENDIX A10 (cont'd.)

Mean Bias and Standard Deviation of Marshall's $\hat{p}_o$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
20	14	.003 / .023	-.001 / .022	-.004 / .023	.021 / .06	-.015 / .029	-.006 / .022	.000 / .02	-.053 / .049
	16	-.001 / .035	.000 / .027	-.009 / .024	-.043 / .043	-.036 / .031	-.017 / .026	-.016 / .026	-.083 / .043
	18	.017 / .05	.016 / .043	-.005 / .033	-.049 / .046	-.037 / .04	-.019 / .043	-.02 / .032	-.013 / .036

APPENDIX A11

Mean Bias and Standard Deviation of Marshall's $\hat{p}_o$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.052 / .042	-.041 / .033	-.081 / .049	.063 / .044	-.045 / .041	-.041 / .032	-.121 / .056	-.018 / .055
10	7	-.021 / .025	-.023 / .023	-.052 / .032	.11 / .04	-.016 / .026	-.024 / .015	-.051 / .025	-.041 / .043
	8	-.015 / .036	-.009 / .029	-.049 / .038	-.032 / .039	-.028 / .028	-.041 / .021	-.051 / .027	-.085 / .032
	9	-.004 / .056	.025 / .036	-.038 / .049	-.077 / .047	-.036 / .045	-.018 / .032	-.012 / .035	-.027 / .05
15	11	-.011 / .025	-.01 / .021	-.032 / .031	.047 / .043	-.018 / .022	-.019 / .019	-.025 / .022	-.042 / .033
	12	-.014 / .03	-.011 / .025	-.029 / .034	-.006 / .043	-.026 / .022	-.027 / .02	-.04 / .024	-.031 / .037
	14	.008 / .042	.025 / .033	.022 / .036	-.035 / .032	-.03 / .052	.011 / .032	.016 / .035	.003 / .038

APPENDIX A11 (cont'd.)

Mean Bias and Standard Deviation of Marshall's $\hat{p}_o$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
20	14	-.006 / .022	-.008 / .019	-.021 / .028	.001 / .041	-.013 / .02	-.01 / .021	-.011 / .021	-.069 / .033
	16	-.009 / .029	-.012 / .026	-.023 / .032	-.038 / .045	-.037 / .025	-.03 / .027	-.031 / .028	-.058 / .042
	18	.012 / .035	.017 / .028	-.002 / .034	-.025 / .03	-.046 / .047	-.021 / .033	-.034 / .033	.01 / .029

APPENDIX A12

Mean Bias and Standard Deviation of Marshall's $\hat{p}_o$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal	Classically Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	-.047 / .031	-.033 / .035	-.076 / .039	.082 / .048	-.045 / .035	-.037 / .037	-.116 / .045	-.036 / .031
10	7	-.02 / .023	-.018 / .027	-.05 / .022	.121 / .036	-.019 / .022	-.02 / .019	-.054 / .021	-.038 / .038
	8	-.009 / .028	-.012 / .032	-.044 / .027	-.047 / .037	-.037 / .035	-.034 / .023	-.054 / .03	-.093 / .032
	9	.005 / .035	.014 / .034	-.034 / .035	-.10 / .035	-.045 / .064	-.015 / .029	-.016 / .04	-.038 / .038
15	11	-.01 / .024	-.007 / .024	-.029 / .021	.049 / .036	-.023 / .022	-.017 / .019	-.026 / .017	-.037 / .033
	12	-.013 / .026	-.011 / .03	-.023 / .023	-.019 / .044	-.036 / .03	-.022 / .022	-.039 / .021	-.035 / .039
	14	.003 / .034	.015 / .034	.027 / .031	-.045 / .025	-.035 / .052	.008 / .029	.011 / .032	-.006 / .028

APPENDIX A12 (cont'd.)

Mean Bias and Standard Deviation of Marshall's $\hat{\rho}_{Q}$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	-.006 / .023	-.002 / .019	-.018 / .02	.016 / .035	-.013 / .02	-.005 / .018	-.013 / .016	-.055 / .033
	16	-.007 / .027	-.01 / .028	-.018 / .021	-.043 / .036	-.042 / .033	-.024 / .027	-.03 / .02	-.064 / .036
	18	.009 / .025	.002 / .035	.006 / .025	-.04 / .026	-.05 / .039	-.028 / .034	-.028 / .025	-.003 / .025

APPENDIX A13

Mean Bias and Standard Deviation of Subkoviak's $\hat{p}_o$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal	Classically Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	-.038 / .046	-.04 / .038	-.07 / .054	.09 / .06	-.05 / .052	-.028 / .048	-.101 / .052	-.009 / .05
10	7	-.013 / .035	-.018 / .028	-.035 / .025	.126 / .055	-.026 / .032	-.019 / .022	-.042 / .023	-.023 / .041
	8	-.006 / .046	-.004 / .031	-.028 / .032	-.039 / .037	-.037 / .045	-.032 / .029	-.04 / .029	-.079 / .044
	9	-.001 / .075	.024 / .034	-.024 / .039	-.101 / .051	-.03 / .064	-.007 / .042	.005 / .037	-.04 / .054
15	11	-.003 / .025	-.006 / .025	-.015 / .025	.05 / .059	-.028 / .031	-.015 / .025	-.015 / .022	-.032 / .047
	12	-.008 / .037	-.006 / .028	-.013 / .027	-.016 / .044	-.036 / .039	-.02 / .03	-.028 / .028	-.038 / .047
	14	.004 / .071	.026 / .045	.02 / .045	-.051 / .051	-.017 / .061	.013 / .041	.035 / .032	-.014 / .038

APPENDIX A13 (cont'd.)

Mean Bias and Standard Deviation of Subkoviak's $\hat{p}_o$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
20	14	.003 / .023	-.001 / .022	-.004 / .023	.021 / .06	-.015 / .029	-.006 / .022	.000 / .02	-.053 / .049
	16	-.001 / .035	.000 / .027	-.009 / .024	-.043 / .043	-.036 / .031	-.017 / .026	-.016 / .026	-.083 / .043
	18	.017 / .05	.016 / .043	-.005 / .033	-.049 / .046	-.037 / .04	-.019 / .043	-.02 / .032	-.013 / .036

APPENDIX A14

Mean Bias and Standard Deviation of Subkoviak's $\hat{p}_o$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.052 / .042	-.041 / .033	-.081 / .049	.063 / .044	-.045 / .041	-.041 / .032	-.121 / .056	-.018 / .055
10	7	-.021 / .025	-.023 / .023	-.052 / .032	.11 / .04	-.016 / .026	-.024 / .015	-.051 / .025	-.041 / .043
	8	-.015 / .036	-.009 / .029	-.049 / .038	-.033 / .039	-.028 / .028	-.041 / .021	-.051 / .027	-.085 / .032
	9	-.004 / .056	.025 / .036	-.038 / .049	-.077 / .047	-.036 / .045	-.018 / .032	-.012 / .035	-.027 / .05
15	11	-.011 / .025	-.01 / .021	-.032 / .031	.047 / .043	-.018 / .022	-.019 / .019	-.025 / .022	-.042 / .033
	12	-.014 / .03	-.011 / .025	-.029 / .034	-.006 / .043	-.026 / .022	-.027 / .02	-.04 / .024	-.031 / .037
	14	.008 / .042	.025 / .033	.022 / .036	-.035 / .032	-.03 / .052	.011 / .032	.016 / .035	.003 / .038

APPENDIX A14 (cont'd.)

Mean Bias and Standard Deviation of Subkoviak's $\hat{p}_{\varrho}$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	-.006 / .022	-.008 / .019	-.021 / .028	.001 / .041	-.013 / .02	-.01 / .021	-.011 / .021	-.069 / .033
	16	-.009 / .029	-.012 / .026	-.023 / .032	-.038 / .045	-.037 / .025	-.03 / .027	-.031 / .028	-.058 / .042
	18	.012 / .035	.017 / .028	-.002 / .034	-.025 / .03	-.046 / .047	-.021 / .033	-.034 / .033	.01 / .029

APPENDIX A15

Mean Bias and Standard Deviation of Subkoviak's $\hat{p}_o$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
5	4	-.047 / .031	-.033 / .035	-.076 / .039	.082 / .048	-.045 / .035	-.037 / .037	-.116 / .045	-.036 / .031
10	7	-.02 / .023	-.018 / .027	-.05 / .022	.121 / .036	-.019 / .022	-.02 / .019	-.054 / .021	-.038 / .038
	8	-.009 / .028	-.012 / .032	-.044 / .027	-.047 / .037	-.037 / .035	-.034 / .023	-.054 / .03	-.093 / .032
	9	.005 / .035	.014 / .034	-.034 / .035	-.10 / .035	-.045 / .064	-.015 / .029	-.016 / .04	-.038 / .038
15	11	-.01 / .024	-.007 / .024	-.029 / .021	.049 / .036	-.023 / .022	-.017 / .019	-.026 / .017	-.037 / .033
	12	-.013 / .026	-.011 / .03	-.023 / .023	-.019 / .044	-.036 / .03	-.022 / .022	-.039 / .021	-.035 / .039
	14	.003 / .034	.015 / .034	.027 / .031	-.045 / .025	-.035 / .052	.008 / .029	.011 / .032	-.006 / .028

APPENDIX A15 (cont'd.)

Mean Bias and Standard Deviation of Subkoviak's $\hat{p}_o$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
20	14	-.006 / .023	-.002 / .019	-.018 / .02	.016 / .035	-.013 / .02	-.005 / .018	-.013 / .016	-.055 / .033
	16	-.007 / .027	-.01 / .028	-.018 / .021	-.043 / .036	-.042 / .033	-.024 / .027	-.03 / .02	-.064 / .036
	18	.009 / .025	.002 / .035	.006 / .025	-.04 / .026	-.05 / .039	-.028 / .034	-.028 / .025	-.003 / .025

APPENDIX A16

Mean Bias and Standard Deviation of Huynh's $\hat{p}_o$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
5	4	-.036 / .04	-.011 / .03	-.026 / .045	.136 / .073	-.047 / .046	-.004 / .037	-.085 / .041	.031 / .056
10	7	-.019 / .027	-.014 / .02	-.049 / .02	.154 / .057	-.031 / .026	-.017 / .016	-.064 / .019	.000 / .04
	8	-.004 / .038	.018 / .021	-.019 / .024	-.004 / .047	-.032 / .04	-.007 / .018	-.045 / .021	-.043 / .049
	9	.03 / .059	.07 / .022	.025 / .026	-.094 / .043	.002 / .058	.046 / .02	.032 / .023	-.031 / .048
15	11	-.012 / .021	.002 / .014	-.026 / .014	.083 / .06	-.04 / .022	-.01 / .013	-.031 / .014	.001 / .049
	12	-.01 / .031	.016 / .015	-.001 / .014	.019 / .052	-.041 / .031	.004 / .014	-.02 / .016	-.006 / .048
	14	.055 / .061	.092 / .017	.09 / .015	-.063 / .045	.041 / .059	.09 / .017	.097 / .017	-.028 / .035

APPENDIX A16 (cont'd.)

Mean Bias and Standard Deviation of Huynh's $\hat{p}_0$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	-.01 / .016	-.002 / .012	-.023 / .011	.046 / .058	-.027 / .017	-.011 / .011	-.025 / .011	-.024 / .054
	16	-.004 / .027	.017 / .013	.001 / .012	-.015 / .047	-.04 / .025	.002 / .013	-.011 / .012	-.064 / .042
	18	.051 / .045	.075 / .015	.061 / .012	-.058 / .042	.003 / .043	.056 / .015	.041 / .013	-.027 / .036

APPENDIX A17

Mean Bias and Standard Deviation of Huynh's $\hat{p}_o$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal	Classically Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	-.048 / .038	-.016 / .023	-.037 / .043	.108 / .053	-.044 / .033	-.012 / .023	-.103 / .044	.016 / .057
10	7	-.025 / .022	-.014 / .015	-.057 / .024	.143 / .04	-.025 / .023	-.021 / .01	-.073 / .021	-.019 / .04
	8	-.008 / .029	.017 / .016	-.03 / .031	.002 / .043	-.029 / .023	-.012 / .011	-.056 / .023	-.052 / .047
	9	.031 / .043	.069 / .017	.013 / .037	-.077 / .04	-.005 / .035	.041 / .013	.02 / .023	-.02 / .042
15	11	-.017 / .018	.001 / .011	-.034 / .02	.084 / .044	-.032 / .02	-.011 / .008	-.039 / .015	-.006 / .034
	12	-.013 / .022	.015 / .011	-.01 / .023	.031 / .046	-.035 / .017	.003 / .008	-.029 / .016	.000 / .035
	14	.058 / .034	.092 / .013	.082 / .026	-.05 / .028	.029 / .044	.09 / .01	.086 / .018	-.013 / .036

APPENDIX A17 (cont'd.)

Mean Bias and Standard Deviation of Huynh's $\hat{p}_o$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	-.016 / .015	-.004 / .01	-.029 / .015	.031 / .041	-.025 / .017	-.012 / .008	-.031 / .012	-.035 / .037
	16	-.009 / .018	.016 / .011	-.006 / .019	-.011 / .045	-.042 / .02	.000 / .009	-.018 / .014	-.044 / .037
	18	.046 / .026	.074 / .011	.053 / .022	-.04 / .028	-.008 / .038	.055 / .01	.032 / .016	-.006 / .029

APPENDIX A18

Mean Bias and Standard Deviation of Huynh's $\hat{p}_o$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.046 .027	-.016 .026	-.036 .029	.129 .062	-.047 .032	-.014 .028	-.099 .035	.000 .04
10	7	-.023 .018	-.015 .018	-.056 .016	.148 .039	-.027 .019	-.023 .014	-.073 .018	-.015 .04
	8	-.005 .024	.016 .019	-.029 .019	-.013 .046	-.036 .031	-.015 .015	-.06 .021	-.06 .046
	9	.036 .03	.067 .02	.013 .023	-.096 .032	-.013 .052	.037 .017	.013 .024	-.032 .034
15	11	-.018 .016	-.001 .014	-.032 .013	.085 .038	-.037 .017	-.015 .01	-.042 .011	-.002 .036
	12	-.015 .021	.013 .014	-.009 .015	.021 .048	-.043 .023	-.001 .01	-.033 .013	-.005 .045
	14	.056 .031	.088 .016	.08 .018	-.06 .023	.024 .049	.083 .012	.08 .018	-.021 .025

APPENDIX A18 (cont'd.)

Mean Bias and Standard Deviation of Huynh's $\hat{p}_o$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal	Classically Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
20	14	-.016	-.005	-.027	.042	-.026	-.014	-.032	-.023
		/.014	/.011	/.01	/.034	/.014	/.009	/.009	/.035
	16	-.011	.014	-.005	-.013	-.046	-.002	-.02	-.049
		/.019	/.012	/.013	/.038	/.023	/.01	/.011	/.035
	18	.044	.071	.052	-.053	-.011	.051	.029	-.018
		/.026	/.013	/.015	/.024	/.039	/.011	/.014	/.024

APPENDIX A19

Mean Bias and Standard Deviation of Subkoviak's $\hat{K}$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.135 / .20	-.089 / .084	-.133 / .127	.007 / .11	-.296 / .187	-.068 / .103	-.306 / .152	-.064 / .065
10	7	-.054 / .199	-.042 / .062	-.088 / .068	.005 / .089	-.212 / .226	-.047 / .052	-.133 / .066	-.082 / .08
	8	.042 / .177	-.013 / .064	-.053 / .078	-.033 / .084	-.137 / .171	-.079 / .068	-.097 / .072	-.179 / .075
	9	.068 / .153	.036 / .07	-.041 / .079	-.026 / .066	-.026 / .143	-.024 / .089	.014 / .08	-.12 / .059
15	11	-.018 / .189	-.021 / .057	-.031 / .058	-.047 / .098	-.197 / .216	-.041 / .057	-.039 / .056	-.086 / .094
	12	-.021 / .17	-.023 / .067	-.023 / .067	.008 / .097	-.13 / .17	-.052 / .07	-.067 / .066	-.11 / .088
	14	.027 / .127	.038 / .089	.03 / .09	-.02 / .067	-.007 / .117	.011 / .087	.064 / .069	-.228 / .049

APPENDIX A19 (cont'd.)

Mean Bias and Standard Deviation of Subkoviak's $\hat{K}$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.05 / .173	-.006 / .05	-.005 / .058	-.076 / .116	-.093 / .19	-.021 / .053	.000 / .054	-.137 / .10
	16	.031 / .136	-.008 / .063	-.011 / .056	-.045 / .114	-.114 / .123	-.047 / .064	-.041 / .071	-.201 / .096
	18	.071 / .098	.02 / .086	-.007 / .071	-.097 / .099	-.058 / .082	-.051 / .087	-.04 / .069	-.192 / .069

APPENDIX A20

Mean Bias and Standard Deviation of Subkoviak's $\hat{K}$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.194 / .192	-.087 / .077	-.143 / .115	-.031 / .057	-.349 / .166	-.087 / .069	-.343 / .141	-.089 / .042
10	7	-.037 / .159	-.048 / .051	-.134 / .11	.005 / .067	-.269 / .189	-.056 / .039	-.159 / .075	-.091 / .065
	8	.054 / .129	-.016 / .062	-.094 / .099	-.037 / .06	-.195 / .158	-.089 / .052	-.122 / .068	-.196 / .055
	9	.082 / .112	.046 / .075	-.064 / .10	-.034 / .049	-.08 / .135	-.039 / .068	-.017 / .072	-.138 / .039
15	11	-.006 / .123	-.023 / .047	-.07 / .087	-.025 / .09	-.263 / .199	-.045 / .046	-.07 / .074	-.095 / .061
	12	-.005 / .096	-.022 / .052	-.048 / .081	.023 / .081	-.19 / .157	-.059 / .049	-.093 / .066	-.124 / .056
	14	.042 / .077	.037 / .074	.029 / .078	-.021 / .055	-.051 / .118	.011 / .069	.03 / .075	-.242 / .036

APPENDIX A20 (cont'd.)

Mean Bias and Standard Deviation of Subkoviak's $\hat{K}$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.043 / .118	-.016 / .044	-.049 / .082	-.087 / .098	-.146 / .151	-.025 / .048	-.035 / .065	-.145 / .065
	16	.034 / .081	-.023 / .054	-.035 / .075	-.064 / .088	-.147 / .109	-.065 / .057	-.07 / .07	-.22 / .062
	18	.071 / .071	.024 / .062	.002 / .069	-.123 / .082	-.082 / .106	-.053 / .071	-.065 / .074	-.212 / .051

APPENDIX A21

Mean Bias and Standard Deviation of Subkoviak's $\hat{K}$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.204 / .115	-.083 / .085	-.144 / .091	-.003 / .087	-.371 / .173	-.091 / .083	-.342 / .132	-.097 / .031
10	7	-.024 / .089	-.052 / .069	-.14 / .077	-.001 / .08	-.28 / .178	-.055 / .059	-.188 / .095	-.104 / .067
	8	.07 / .06	-.034 / .072	-.091 / .074	-.041 / .071	-.20 / .143	-.083 / .055	-.138 / .091	-.20 / .062
	9	.092 / .052	.021 / .072	-.057 / .073	-.035 / .053	-.079 / .13	-.04 / .055	-.027 / .086	-.137 / .047
15	11	.002 / .10	-.027 / .061	-.073 / .07	-.036 / .09	-.248 / .173	-.051 / .06	-.083 / .071	-.085 / .075
	12	-.004 / .085	-.032 / .066	-.042 / .061	.014 / .085	-.188 / .134	-.056 / .057	-.10 / .072	-.109 / .07
	14	.034 / .062	.023 / .07	.047 / .064	-.024 / .054	-.046 / .095	.008 / .057	.021 / .065	-.23 / .037

APPENDIX A21 (cont'd.)

Mean Bias and Standard Deviation of Subkoviak's $\hat{K}$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.045 / .097	-.013 / .053	-.051 / .071	-.068 / .089	-.135 / .161	-.023 / .056	-.049 / .07	-.119 / .073
	16	.036 / .081	-.027 / .062	-.031 / .054	-.045 / .076	-.159 / .081	-.061 / .061	-.075 / .06	-.187 / .067
	18	.062 / .056	-.005 / .072	.018 / .052	-.11 / .059	-.088 / .017	-.064 / .066	-.053 / .052	-.192 / .043

APPENDIX A22

Mean Bias and Standard Deviation of Huynh's $\hat{K}$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	Randomly Parallel Alternate Forms: J-shaped	Randomly Parallel Alternate Forms: Bi-modal	Randomly Parallel Alternate Forms: Normal	Classically Parallel Alternate Forms: Skewed	Classically Parallel Alternate Forms: J-shaped	Classically Parallel Alternate Forms: Bi-modal	Classically Parallel Alternate Forms: Normal
5	4	-.063 / .168	-.024 / .065	-.03 / .109	.107 / .145	-.193 / .172	-.012 / .077	-.249 / .113	.033 / .102
10	7	-.01 / .141	-.03 / .044	-.114 / .055	.077 / .108	-.132 / .167	-.04 / .04	-.197 / .05	-.013 / .096
	8	.088 / .127	.035 / .044	-.025 / .051	.046 / .107	-.061 / .15	-.018 / .04	-.105 / .045	-.101 / .092
	9	.147 / .117	.133 / .046	.059 / .05	.053 / .098	.067 / .136	.089 / .042	.073 / .042	-.039 / .083
15	11	.000 / .137	.002 / .032	-.043 / .032	.03 / .114	-.18 / .148	-.025 / .031	-.07 / .035	-.013 / .10
	12	.016 / .125	.031 / .032	.018 / .03	.087 / .113	-.088 / .136	.006 / .031	-.032 / .033	-.031 / .098
	14	.134 / .112	.18 / .035	.174 / .03	.056 / .10	.122 / .116	.173 / .034	.191 / .032	-.147 / .083

APPENDIX A22 (cont'd.)

Mean Bias and Standard Deviation of Huynh's $\hat{K}$ Across Samples of 25 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.037 / .116	-.005 / .029	-.038 / .027	-.01 / .123	-.106 / .116	-.025 / .028	-.055 / .031	-.073 / .113
	16	.064 / .099	.034 / .029	.024 / .025	.028 / .123	-.082 / .099	.001 / .028	-.011 / .028	-.129 / .11
	18	.155 / .089	.143 / .031	.128 / .025	-.023 / .116	.04 / .087	.106 / .029	.088 / .027	-.121 / .093

APPENDIX A23

Mean Bias and Standard Deviation of Huynh's $\hat{K}$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
5	4	-.111 / .18	-.032 / .053	-.043 / .10	.064 / .091	-.248 / .169	-.023 / .049	-.281 / .107	-.003 / .078
10	7	.012 / .115	-.026 / .032	-.134 / .084	.083 / .082	-.194 / .146	-.045 / .024	-.216 / .05	-.032 / .09
	8	.107 / .102	.039 / .032	-.044 / .078	.047 / .079	-.117 / .133	-.024 / .024	-.124 / .046	-.122 / .088
	9	.164 / .092	.137 / .034	.039 / .076	.048 / .073	.018 / .118	.082 / .026	.053 / .044	-.06 / .077
15	11	.014 / .084	.004 / .023	-.059 / .058	.056 / .094	-.239 / .148	-.022 / .017	-.088 / .043	-.018 / .068
	12	.028 / .076	.033 / .023	.001 / .055	.11 / .092	-.142 / .134	.009 / .017	-.05 / .04	-.039 / .066
	14	.144 / .065	.181 / .026	.155 / .056	.065 / .083	.081 / .109	.175 / .02	.172 / .038	-.163 / .056

APPENDIX A23 (cont'd.)

Mean Bias and Standard Deviation of Huynh's $\hat{K}$ Across Samples of 35 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.034 / .073	-.002 / .021	-.052 / .049	-.014 / .098	-.147 / .12	-.024 / .018	-.069 / .04	-.074 / .076
	16	.06 / .062	.036 / .021	.01 / .045	.017 / .094	-.117 / .101	.002 / .018	-.025 / .036	-.14 / .073
	18	.15 / .054	.144 / .023	.114 / .045	-.046 / .089	.011 / .086	.106 / .02	.072 / .034	-.142 / .062

APPENDIX A24

Mean Bias and Standard Deviation of Huynh's $\hat{K}$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal	Classically Parallel Alternate Forms: Skewed	J-shaped	Bi-modal	Normal
5	4	-.112 / .104	-.041 / .061	-.052 / .074	.099 / .12	-.274 / .164	-.038 / .062	-.28 / .10	-.012 / .061
10	7	.017 / .056	-.036 / .042	-.139 / .054	.067 / .097	-.195 / .131	-.057 / .034	-.237 / .063	-.04 / .094
	8	.112 / .05	.03 / .041	-.048 / .049	.035 / .094	-.117 / .118	-.036 / .034	-.142 / .056	-.128 / .093
	9	.169 / .046	.129 / .042	.036 / .047	.041 / .086	.018 / .107	.07 / .036	.037 / .051	-.063 / .084
15	11	.013 / .067	-.005 / .032	-.064 / .041	.045 / .098	-.234 / .118	-.036 / .025	-.104 / .045	-.012 / .09
	12	.027 / .061	.024 / .032	-.002 / .038	.101 / .095	-.137 / .108	-.005 / .025	-.064 / .042	-.031 / .087
	14	.143 / .054	.173 / .035	.154 / .037	.061 / .081	.083 / .092	.161 / .026	.159 / .037	-.152 / .068

APPENDIX A24 (cont'd.)

Mean Bias and Standard Deviation of Huynh's $\hat{K}$ Across Samples of 50 Examinees

Test Length	Cut-off Score	Randomly Parallel Alternate Forms				Classically Parallel Alternate Forms			
		Skewed	J-shaped	Bi-modal	Normal	Skewed	J-shaped	Bi-modal	Normal
20	14	.03 / .06	-.011 / .027	-.056 / .035	.001 / .088	-.148 / .098	-.034 / .023	-.08 / .038	-.051 / .082
	16	.056 / .051	.028 / .027	.007 / .032	.035 / .085	-.119 / .083	-.007 / .023	-.035 / .034	-.115 / .079
	18	.147 / .046	.137 / .028	.112 / .031	-.026 / .078	.008 / .074	.098 / .023	.064 / .031	-.117 / .067

LIST OF REFERENCES

LIST OF REFERENCES

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

Anastasi, A. Psychological testing. New York: Macmillan, 1976.

Berk, R. A. Item analysis. In R. Berk (Ed.), Criterion-referenced testing: State of the art. Baltimore: The Johns Hopkins University Press, 1980.

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

Block, J. H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-190.

Brennan, R. L. Psychometric methods for criterion-referenced tests. Unpublished manuscript, March 1974. (Available from [Department of Education, SUNY at Stony Brook, Stony Brook, New York, 11790]).

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests (ACT Technical Bulletin No. 27). Iowa City, Iowa: American College Testing Program, December 1977.

Brennan, R. L. Extensions of generalizability theory to domain-referenced testing (ACT Technical Bulletin No. 30). Iowa City, Iowa: American College Testing Program, June 1978.

Brennan, R. L. Some applications of generalizability theory to the dependability of domain-referenced tests (ACT Technical Bulletin No. 32). Iowa City, Iowa: American College Testing Program, April 1979.

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

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, 609-625. (b)

Brennan, R. L., & Lockwood, R. E. A comparison of two cutting score procedures using generalizability theory (ACT Technical Bulletin No. 33). Iowa City, Iowa: American College Testing Program, April 1979.

Buck, L. S. Use of criterion-referenced tests in personnel selection: A summary status report (Technical Memorandum 75-6). Washington, D.C.: United States Civil Service Commission, December 1975.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative Research: Strategies and Methods. Washington, D.C.: American Institutes for Research, 1970.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Downing, S. M., & Mehrens, W. A. Six single-administration reliability coefficients for criterion-referenced tests: A comparative study. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Ebel, R. L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Eignor, D. R., & Hambleton, R. K. Relationship of test length to criterion-referenced test reliability and validity (Report No. 86). Amherst: University of Massachusetts (School of Education), Laboratory of Psychometric and Evaluative Research, 1979.

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 system development. New York: Holt, 1962.

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.

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

Goldstein, I. L. Training program: Development and evaluation. Monterey, California: Brooks/Cole, 1974.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49, 732-764.

Gross, A. L., & Schulman, V. The applicability of the beta binomial model for criterion referenced testing. Journal of Educational Measurement, 1980, 17, 195-201.

Hambleton, R. K., & Eignor, D. R. Criterion-referenced test development and validation methods. Training program presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

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

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

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

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

Harris, M. L., & Stewart, D. M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

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

Huynh, H., & Saunders, J. C., III. Accuracy of two procedures for estimating reliability of mastery tests. Paper presented at the annual conference of the Eastern Educational Research Association, Kiawah Island, South Carolina, February 1979.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, August 1970.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests (ACT Technical Bulletin No. 28). Iowa City, Iowa: American College Testing Program, December 1977.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Klein, S. P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests (ERIC/TM Report 26). Princeton: ERIC Clearinghouse on Tests, Measurement, and Evaluation, September 1973.

Koslowsky, M., & Bailit, H. A measure of reliability using qualitative data. Educational and Psychological Measurement, 1975, 35, 843-846.

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

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26. (b)

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

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

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

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

Lovett, H. T. Criterion-referenced reliability estimated by ANOVA. Educational and Psychological Measurement, 1977, 37, 21-29.

Magnusson, D. Test theory. Reading, Massachusetts: Addison-Wesley, 1967.

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

Marshall, J. L. Possible mathematical relationships of true and obtained scores and their implications for mastery testing. Paper presented at the annual meeting of the Midwest Educational Research Association, Bloomingdale, Illinois, 1978.

Marshall, J. L. Personal communication, 1980.

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., March-April 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 annual meeting of the American Educational Research Association, San Francisco, April 1979.

Mehrens, W. A., & Ebel, R. L. Some comments on criterion-referenced and norm-referenced achievement tests. Measurement in Education, Winter 1979, 10, 1-8.

Michigan Department of Education. Technical Report: Michigan Educational Assessment Program. Lansing: Michigan Department of Education, Research, Evaluation and Assessment Services, June 1977.

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

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

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3, Los Angeles: Center for the Study of Education, University of California, 1974.

Peng, C.-Y. J., An investigation of Huynh's normal approximation procedure for estimating criterion-referenced reliability. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Peng, C.-Y. 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., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Schmitt, N., & Schmitt, K. Differences in reliability estimates for objective-referenced tests. Unpublished manuscript, 1977.

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

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. Further comments on reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1977.

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

Subkoviak, M. J. The reliability of mastery classification decisions. Paper presented at the first annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., 1978. (b).

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

Wardrop, J. L., Anderson, T. H., Hively, W., Hastings, C. N., Anderson, R. I., Muller, K. E. A framework for analyzing the inference structure of educational achievement tests. Journal of Educational Measurement, 1982, 19, 1-18.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64. (a)

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140. (b)

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03175 1898