THE CHARACTERISTICS AND PROPERTIES OF THE THRESHOLD AND SQUARED-ERROR CRITERION-REFERENCED AGREEMENT INDICES

DISSERTATION

Presented to the Graduate Council of the North Texas State University in Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Cynthia F. Dutschke, B.S., M.A.

Denton, Texas

May, 1988
Educators who use criterion-referenced measurement to ascertain the current level of performance of an examinee in order that the examinee may be classified as either a master or a nonmaster need to know the accuracy and consistency of their decisions regarding assignment of mastery states.

This study examined the sampling distribution characteristics of two reliability indices that use the squared-error agreement function: Livingston's $\frac{k^2(x,T_x)}{x}$ and Brennan and Kane's $M(C)$. The sampling distribution characteristics of five indices that use the threshold agreement function were also examined: Subkoviak's $P_C$, Huynh's $p$ and $k$, and Swaminathan's $p$ and $k$. These seven methods of calculating reliability were also compared under varying conditions of sample size, test length, and criterion or cutoff score.

Computer-generated data provided randomly parallel test forms for $N = 2000$ cases. From this, 1000 samples were drawn, with replacement, and each of the seven reliability
indices was calculated. Descriptive statistics were collected for each sample set and examined for distribution characteristics. In addition, the mean value for each index was compared to the population parameter value of consistent mastery/nonmastery classifications.

The results indicated that the sampling distribution characteristics of all seven reliability indices approach normal characteristics with increased sample size. The results also indicated that Huynh's $p$ was the most accurate estimate of the population parameter, with the smallest degree of negative bias. Swaminathan's $p$ was the next best estimate of the population parameter, but it has the disadvantage of requiring two test administrations, while Huynh's $p$ index only requires one administration.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................ iv

Chapter

I. INTRODUCTION ................................................................. 1

  Problem
  Research Questions
  Significance of the Study
  Assumptions

II. REVIEW OF THE LITERATURE ........................................... 9

  Squared-Error Loss Function
    Livingston's $k^2(x, I_n)$
    Brennan and Kane's $-M(C)$

  Threshold Loss Function
    Swaminathan's $p$ and $k$
    Huynh's $p$ and $k$
    Subkoviak's $p$

  Previous Comparative Studies
  Use of an Indifference Zone
  Losses Due to Misclassification
  Degree of Mastery or Nonmastery

III. METHOD OF RESEARCH ............................................... 34

  Description of Simulation

IV. PRESENTATION AND ANALYSIS OF DATA ....................... 41

  Analysis
    Research Questions 1a-5

V. FINDINGS, CONCLUSIONS, RECOMMENDATIONS .............. 79

  Findings
  Conclusions
  Recommendations

APPENDICES ........................................................................ 90

BIBLIOGRAPHY ..................................................................... 114
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Descriptive Statistics for Two Populations</td>
<td>42</td>
</tr>
<tr>
<td>2. Descriptive Statistics for Livingston's $k^2(X,T_x)$ Normally-skewed Population</td>
<td>43</td>
</tr>
<tr>
<td>3. Descriptive Statistics for Livingston's $k^2(X,T_x)$ Negatively-skewed Population</td>
<td>45</td>
</tr>
<tr>
<td>4. Descriptive Statistics for Brennan and Kane's $M(C)$ Normally-skewed Population</td>
<td>47</td>
</tr>
<tr>
<td>5. Descriptive Statistics for Brennan and Kane's $M(C)$ Negatively-skewed Population</td>
<td>48</td>
</tr>
<tr>
<td>7. Descriptive Statistics for Huynh's $p$ Negatively-skewed Population</td>
<td>52</td>
</tr>
<tr>
<td>8. Descriptive Statistics for Huynh's $k$ Normally-skewed Population</td>
<td>54</td>
</tr>
<tr>
<td>14. Descriptive Statistics for Swaminathan's $k$ Normally-skewed Population</td>
<td>64</td>
</tr>
<tr>
<td>15. Descriptive Statistics for Swaminathan's $k$ Negatively-skewed Population</td>
<td>66</td>
</tr>
</tbody>
</table>
LIST OF TABLES (continued)

16. Means of Reliability Estimates
   Normally Distributed Population (N=30) .... 68

17. Means of Reliability Estimates
   Normally Distributed Population (N=150) .... 70

18. Means of Reliability Estimates
   Negatively-skewed Population (N=30) .... 72

19. Means of Reliability Estimates
   Negatively-skewed Population (N=150) .... 74

20. Cutoff Score Conditions of Estimated Reliability
    Values (N=150, n=10) ................. 76
CHAPTER 1

INTRODUCTION

A criterion-referenced test is defined by Popham (1981) as one used to "ascertain an individual's status (referred to as a domain score) with respect to a well-defined behavior" (p.130). The concept of domain is fundamental to the definition of criterion-referenced measurement. This concept is critical to appreciating the importance of reliability estimation for these tests.

Any universe of behaviors can be partitioned into a potentially infinite number of domains. Each domain is defined in such a manner that a set of behaviors is either included in it or excluded from it. The criterion-referenced test, then, is one that is deliberately constructed to produce measurements that are interpreted directly in terms of specific performance standards which are usually specified by defining some domain of tasks that the student should perform. Samples of tasks that are representative of his domain are organized into a test. Measurements are then taken and are used to make a decision about the performance of each individual relative to that particular domain (Glaser & Nitko, 1971). Many tests measure more than one objective and are therefore arranged into subtests that do not overlap,
but are related directly to the objectives measured by the test (Hambleton, Swaminathan, Algina & Coulson, 1978). An instructional decision for each individual is then often made on the basis of performance of each subtest.

Glaser (1963) and Popham and Husek (1969) were the first to introduce the use of criterion-referenced testing, with the goal of providing the kind of test score information needed to make the many decisions necessitated by the rise in objectives-based programs. Norm-referenced tests were viewed as inappropriate for providing this kind of information (Popham, 1981).

The popularity of criterion-referenced tests today is in evidence everywhere. Criterion-referenced tests are used to monitor student progress in objectives-based instruction, to diagnose deficiencies in learning, to carry-out program evaluations, as well as to assess competency on numerous certification and licensing examinations (Hambleton, 1974).

According to Subkoviak (1978a, p. 268)) a mastery test is defined as "a test with a single cutting score, \( c \), that determines mastery and nonmastery classes - scores above and below \( c \), respectively." With the increased popularity of criterion-referenced tests, the users are constantly faced with decisions about mastery-nonmastery classifications that are made from the results of these tests. If such decisions are made that classify examinees accordingly, then knowledge of the reliability of these decisions becomes extremely
important to the test user who must accept the decisions and to the examinees who must perceive the results as appropriate (Reid, 1984).

For example, if decisions concerning one's mastery states on the objectives of a test are used to monitor the examinee's progress through the program, then the examiner must assign the examinee to the state of either mastery or nonmastery in such a way to minimize the chosen loss function (Lord & Novick, 1973). The threshold of loss function treats all errors of classification as equivalent. It focuses upon error as it has an impact on the qualitative judgment involved in assigning a person to a mastery level. In this loss function, there is no distinction between the seriousness of a false-positive or a false-negative decision. The squared-error loss function, on the other hand, is a random-effects linear model that gives equal weight to errors having an impact on the ordering of scores throughout the entire range of scores. It focuses upon the error involved in all quantitative judgments and treats some classification errors as more serious than others. Indeed, if on the basis of a student's performance on a sample of items measuring a specific objective, the decision is made to advance the student, retain him temporarily, or retain him for an indefinite period of time, these decisions must be reliable.

A comprehensive study of the characteristics of the loss functions with regard to varying conditions of sample size,
sample distribution, and cut-off score placement would be very beneficial to the many users of criterion-referenced tests.

Problem

Educators regularly find themselves faced with the problem of ascertaining the current level of performance of an examinee, with regard to the course content, in order that the examinee may be classified as either a master or a nonmaster. Those who use criterion-referenced measurement to assess this level of performance and/or achievement need to know the accuracy and consistency of their decisions regarding mastery status.

Research Questions

To fulfill the purpose of this study, the following research questions were answered:

1a. What are the sampling distribution characteristics of Livingston's $k^2(X, T_x)$, a squared-error loss function, as used in determining the reliability of criterion-referenced tests?

b. What are the sampling distribution characteristics of Brennan and Kane's $M(C)$, a squared-error loss function, as used in determining the reliability of criterion-referenced tests?
2a. What are the sampling distribution characteristics of Huynh's $p$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

b. What are the sampling distribution characteristics of Huynh's $k$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

c. What are the sampling distribution characteristics of Subkoviak's $P_E$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

d. What are the sampling distribution characteristics of Swaminathan's $p$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

e. What are the sampling distribution characteristics of Swaminathan's $k$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

3. How do the methods of Livingston, Brennan and Kane, Huynh, Subkoviak, and Swaminathan compare with respect to their estimation of the proportion of consistent classifications (reliability) of the population?

4. How are the indices of reliability which use the threshold loss and squared-error loss function compatible with the condition of preferring to use two cutoff scores with a zone of indifference in the middle?

5. How are the different reliability indices, with regard to either squared-error loss or threshold loss
functions, compatible under conditions where losses related to all false mastery and false nonmastery decision errors are not equally serious?

Significance of the Study

The study compared the reliability indices of criterion-referenced tests, determined the sampling distribution characteristics of each index, and determined under what conditions these indices were compatible with specific conditions of cutoff scores.

Comparative studies have been done in the past, but according to Berk (1984) there is a need for research investigating compatibility with various cutoff score conditions. In addition, Subkoviak (1978a) states there is a need for additional studies comparing the characteristics and properties of various methods of establishing reliability.

This study is significant in that it states conditions under which squared-error loss or threshold loss agreement indices should be used and it clarifies the compatibility of these indices with varying conditions of criterion, or cutoff scores.

Assumptions

The following assumptions were made regarding the data that was randomly generated by computer for this study:

1. Local independence - for a fixed examinee, the
scores on items are statistically independent.

2. Equivalent measures - for a fixed examination, the scores on separate tests are independently and identically distributed.


CHAPTER 2

REVIEW OF THE LITERATURE

As has been pointed out by many authors, particularly by Peng and Subkoviak (1980), the appropriate techniques for empirical estimation of criterion-referenced reliability are still undecided. It is essential, however, in criterion-referenced measures that a test be reliable. If a test is found to be unreliable, it is useless to establish its validity, since unreliability, alone, renders a test inappropriate for use. In determining the reliability of a criterion-referenced test, the issue generally involves the consistency of mastery/nonmastery classifications over repeated measurements. The criterion-referenced reliability coefficient gives the test user a group of scores relative to a cutoff score. It provides data about what proportion of the information provided by a test is reliable information (Livingston, 1972).

Reliability indices used with norm-referenced tests are inappropriate for criterion-referenced tests because of the difficulties that arise when a correlation coefficient is applied (Millman, 1974). In mastery learning situations, it is very possible for a criterion-referenced test score
distribution to have little or no variance and yet the test scores may be very reliable. A correlation coefficient that requires variance in the distribution, such as that in norm-referenced reliability indices, would not reflect this reliability. In fact, in criterion-referenced tests where little variance is often found, zero and even negative correlation coefficients may be calculated if the same index is used to calculate criterion-referenced reliability that is used to calculate norm-referenced reliability.

Criterion-referenced test scores are theoretically intended to have meaning that is not dependent on the performance of other examinees who take the test, but is used to obtain an absolute meaning of achievement with respect to a well-defined domain of behaviors (Lang, 1980). To determine the reliability of these tests, three indices are defined by Hambleton et al. (1978): (1) reliability of mastery classification decisions or consistency of mastery-nonmastery decisions across repeated testings with either one test or parallel test forms, referred to as threshold loss function; (2) reliability of criterion-referenced test scores or consistency of squared deviations of individual scores from the cutoff score across parallel test forms, referred to as squared-error loss function; and (3) reliability of domain score estimates or consistency of individual scores across parallel test forms. These indices are appropriate when a cutoff score is the basis of the classification decision.
The squared-error loss function deals with the consistency of measurements or scores and is based on the "squared deviations of individual scores from the cut-off score" (Berk, 1984). Because of this basis, it employs a sensitivity to degree of mastery and nonmastery along score continuum as well as the qualitative master-nonmaster classification assumed by this function. In addition, the losses associated with false positive and false negative decisions are not considered equally serious, as they are in threshold loss function. This function considers misclassifying examinees far away from the cut-off score more serious than misclassification of those close to that score. This sensitivity to degree of mastery, however, also means that this index indicates the magnitude of all errors of measurement, including those that do not lead to misclassification. In other words, it gives equal weight to errors having an impact on the ordering of scores throughout the entire range of scores. It focuses upon the error involved in all quantitative judgments. This is viewed as a fundamental problem with this loss function (Brennan and Kane, 1977b).

\[ \text{Livingston's } k^2(X, T_x) \]

Livingston (1972), as one of the pioneers in criterion-referenced test reliability, demonstrates how classical test
theory can be used to derive a reliability coefficient for criterion-referenced measures that parallels that of norm-referenced measures. He asserts that the basic difference between norm-referenced and criterion-referenced tests is that in norm-referenced one is concerned with the deviation of the score from the group mean and in criterion-referenced tests one is concerned with the deviation from a fixed standard: the criterion or cutoff score.

Substituting the criterion score for the mean of the norm group, Livingston then defines the various indices accordingly. Variance, for example, is replaced with his "mean squared deviation" \( D^2(X) \) by calculating the deviation of individual scores \( X \) about the criterion score \( C_x \).

The expected value \( (\epsilon) \) becomes

\[
D^2 (X) = \epsilon (X - C_x)^2
\]

and

\[
D^2 (X) = \sigma^2 (X) + (\mu_x - C_x)^2
\]

and for true scores

\[
D^2 (T_x) = \alpha_{21} (X, T_x) \sigma^2 (X) + (\mu_x - C_x)^2
\]

where \( \alpha_{21} \) is the Kuder-Richardson reliability coefficient.

When the mean and the criterion are identical, classical reliability will equal criterion-referenced reliability; otherwise, it is always greater than classical reliability (Shavelson et al., 1972). Livingston (1972) then states that
since the mean squared deviation of obtained scores and of true scores can be expressed in terms of norm-referenced indices, that the same can be done for their ratio, which is the criterion-referenced reliability coefficient:

\[
k^2(X,T_x) = \frac{D^2(T_x)}{D^2(X)} = \frac{\alpha_2(X,T_x)}{\sigma^2(X)} \frac{\sigma^2(X) + (\mu_X - C_X)^2}{\sigma^2(X) + (\mu_X - C_X)^2}
\]

where \(\alpha_2(X,T_x)\) is the norm-referenced reliability coefficient.

**Brennan and Kane's M(C)**

Brennan and Kane (1977a) develop an index to measure the dependability of mastery-nonmastery decisions based on the testing procedure. In the treatment of error variance for domain-referenced measures, Brennan and Kane note that the differences one wishes to find are in the form of \(\mu - c\) and the observed differences are of the form \(X_{pi} - c\), where \(c\) is a mastery cut-off score, \(X_{pi}\) the raw score, and \(\mu_p\) the mean for person \(p\) (Brennan, Berk, ed., 1980). An index of dependability is then defined with \(c\) replacing \(\mu\), in terms of expected squared deviation from \(c\), rather than in terms of variances, as:

\[
M(C) = \frac{\epsilon_p (\mu_p - C)^2}{\epsilon_i \epsilon_p (X_{pi} - C)^2}
\]

If items are scored dichotomously, an estimation of \(M(C)\) is relatively straightforward and can be expressed as:

\[
M(C) = 1 - \left( \frac{1}{np-1} \sum_{i=1}^{np} \frac{X_{pi} (1-X_{pi}) - S^2(X_{pi})}{(X_{pi} - C)^2 + S^2(X_{pi})} \right)
\]
where $\frac{s^2}{\sum p_i}$ is the sample variance of a person's mean scores over items.

The mathematically intricate derivation of Equation 6 relies on the fact that the numerator and denominator of Equation 5 are expected squared deviations rather than variances. The derivation, then, is similar to Livingston's derivation of his coefficient (Brennan and Kane, 1977a). The main difference is that the derivation of Brennan and Kane's $M(C)$ is based on the assumption of randomly parallel tests and Livingston's is based on classically parallel tests (Brennan and Kane, 1980).

Both Livingston's and Brennan and Kane's coefficients assume a squared-error loss function while the indices to be examined in the next section are based on the threshold loss, which implies that "only misclassifications of true masters and nonmasters constitute errors, and that all such misclassifications are equally serious. For a mastery test, it is not clear that all misclassifications are equally serious" (Brennan and Kane, 1977a, p. 286). Brennan and Kane feel that it is frequently impossible to make a clear distinction between masters and nonmasters and that cut-off scores are inevitably somewhat arbitrary.

Whereas neither of these loss functions is ideal, a squared-error loss function has the advantage of being sensitive to many errors, but the disadvantage of being sensitive to all errors of measurement, including those not
leading to misclassification (Brennan and Kane, 1977a).

**Threshold Loss Function**

Hambleton and Novick (1973) were the first to suggest the use of a threshold loss function to determine classification decision consistency. The threshold loss function assumes that "a dichotomous qualitative classification of students as masters and nonmasters of an objective based on a threshold or cut-off scores exists and the losses associated with all false mastery and false nonmastery classification errors are equally serious regardless of their size" (Berk, 1984).

Hambleton and Novick (1973) explain threshold loss in terms of some criterion level $\pi_0$ and define a parameter $\omega$ such that

\[
\omega = 1 \text{ if } \pi \geq \pi_0
\]
\[
\omega = 0 \text{ if } \pi < \pi_0
\]

Examinees with $\omega$ values of 1 are those who have true ability levels equal to or greater than the criterion level $\pi_0$, and those having values of zero are those whose true ability levels are below $\pi_0$. If an estimate of $\hat{\pi}_i$ can be obtained, then an estimate of $\hat{\omega}$ could be obtained as follows:

\[
\hat{\omega} = 1 \text{ if } \hat{\pi}_i \geq \pi_0
\]
\[
\hat{\omega} = 0 \text{ if } \hat{\pi}_i < \pi_0
\]
Defining the error of estimation as \((\hat{\omega} - \omega)\), the difference between the estimated and true value, it takes on three values, \(+1\), \(-1\), or \(0\), corresponding to a false positive error, a false negative error, or a correct classification. The criterion-referenced measurement problem, then, becomes one of calculating an estimator \(\hat{\omega}\) of \(\omega\) by determining an estimator \(\hat{\pi}_1\) or \(\pi_0\) with a threshold loss function and converting this to an estimate of \(\pi\).

Hambleton and Novick (1973) further suggest that the reliability of mastery classification decisions should be defined in terms of the consistency of decisions from two administrations of the same test or parallel forms of a test. Their index, then, is

\[
P_o = \sum_{k=1}^{m} p_{kk}
\]

(7)

where \(p_{kk}\) is the proportion of examinees classified in the \(k\)th mastery level on the two administrations. The index \(p_o\) then is the observed proportion of decisions that are in agreement.

Swaminathan's \(p\) and \(k\)

Swaminathan, Hambleton, and Algina (1974) extended the work of Hambleton and Novick by administering a criterion-referenced test to a population of examinees on two occasions and then classifying the examinees into one of \(k\) mutually exclusive mastery states (master/nonmaster) at each occasion.
administration, \( P_{ii} \). This represents the proportion of examinees placed at the \( i \)th mastery state on the first administration and at the \( j \)th mastery state on the second administration. The measure of agreement between the two decisions is then represented by the formula

\[
P_0 = \sum_{i=1}^{k} P_{ii}
\]

This represents the proportion of examinees placed in the \( i \)th mastery state on both administrations. Since this index of agreement does not take agreement expected by chance into account, Swaminathan et al. (1974) choose to use the kappa, \( k \), coefficient introduced by Cohen (1960) as a coefficient of agreement between the nominal scales. Using coefficient \( k \), the expression for reliability of criterion-tests becomes

\[
k = \frac{P_0 - P_c}{1 - P_c}
\]

where \( P_0 \), the observed proportion of agreement, is given by equation 8 above. The expected proportion of agreement, \( P_c \), is given by

\[
P_c = \sum_{i=1}^{k} P_{i}. P_{i}
\]

where the proportions of examinees assigned to the mastery state \( i \) on the first and second test administrations are denoted by \( P_{i}. \) and \( P_{i} \) respectively.

Coefficient kappa, then, can be considered as the proportion of examinees consistently classified as masters.
and nonmasters beyond that expected by chance alone. It represents the proportion by which test scores improve the consistency of decisions beyond that expected by chance. The upper limit of $k$ is +1. However, if any examinee is classified differently on repeated examinations, the value of $k$ will always be less than +1. The lower limit of $k$, according to Swaminathan, is around -1. Preciseness of the lower limit is of little concern, since any negative value of $k$ indicates classification inconsistency and, therefore, unreliability of the test in question. To estimate the value of $k$ from only a sample of examinees, rather than the entire population of examinees, Swaminathan et al (1974) defined the estimate $\hat{k}$ of $k$ as the sample analog of equation (7), by substituting the sample frequencies and proportions for the population frequencies and proportions.

**Huynh's $p$ and $k$**

Huynh (1976b) describes a model for the evaluation of the kappa reliability index on the basis of a single test administration, since practical situations, he feels, often make repeated testings impossible and it therefore is desirable to have a method of estimating the reliability of decisions on the basis on a single administration. He makes use of the beta-binomial model, stating that it is "particularly suited to scores obtained from a domain-referenced test" (p. 253). Indeed, Gross and Shulman (1980) found that the beta-binomial model is robust to violations
in its underlying assumptions, and is therefore suitable for use in criterion-referenced testing situations, where these assumptions frequently do not hold. There are also other indications that the model adequately fits many test score distributions (Huynh and Saunders, 1979).

Using the beta-binomial model, which automatically assumes that all items in the universe are exchangeable, and assuming local independence between two equivalent test forms X and Y, both of length n, the bivariate distribution \( f(X, Y) \) can be derived as a bivariate negative hypergeometric (or beta-binomial) density with parameters \( \alpha \) and \( \beta \):

\[
f(x) = \binom{n}{x} B(\alpha + x, n + \beta - x) / B(\alpha, \beta) \quad (11)
\]

\[
f(x, y) = \frac{\binom{n}{x} \binom{n}{y}}{B(\alpha + x + y, 2n + \beta - x - y)} B(\alpha + x + y, 2n + \beta - x - y) \quad (12)
\]

Using single estimates from a single test administration for the parameters \( \alpha \) and \( \beta \), Huynh (1976a) proposed to estimate the proportions in the matrix cells, and after that, to apply the usual procedure for computing coefficient kappa, by first giving the following proportions:

\[
P_{ii} = \sum_{x, y = c}^{n} f(x, y) \quad (13)
\]

and

\[
P_{i} = \sum_{x = c}^{n} f(x) \quad (14)
\]

where \( P_{ii} \) is the proportion of examinees granted mastery status on both \( X \) and \( Y \) test forms, and \( P_{i} \) is the proportion granted mastery status on the basis of data obtained from
either test form. Through algebraic manipulation, the following formula for the reliability index kappa is obtained by Huynh (1976b):

\[
k = \frac{(P_{ii} - P_i)^2}{(P_i - P_i^2)}
\]  

(14)

which is most useful with cutoff scores are near \( n \). If there are small cutoff scores, the following formula is more efficient:

\[
k = \frac{(P_{oo} - P_o)^2}{(P_o - P_o^2)}
\]  

(15)

where \( P_{oo} \) and \( P_o \) are the proportions of examinees not granted mastery status by one or by both forms, respectively. In other words,

\[
P_{oo} = \sum_{x=0}^{c-1} f(x,y)
\]  

(16)

and

\[
P_o = \sum_{x=0}^{c-1} f(x)
\]  

(17)

If a theoretical look at kappa, Huynh (1976b) proves that this index cannot have negative values, contrary to Swaminathan's statement that -1 is the lower limit for \( k \), when it is based on data obtained from two equivalent forms, which in turn indicates the problems which might arise in estimating this index on the test-retest scores of a small group of examinees.
Huynh's method has been very popular perhaps because of its requirement of only one test administration and because it produces reliability estimates involving little bias or error (Peng and Subkoviak, 1980). Because of its use of the beta-binomial distribution, however, it is very difficult to calculate.

Because of its mathematical complexity, Peng and Subkoviak (1980) have derived a normal approximation of Huynh's method of computing the $p$ and $k$ index of reliability. The first step is to obtain an estimated value of $z$ by use of the following formula:

$$
\hat{z} = \frac{c - .5 - \mu}{\sigma} \tag{18}
$$

where .5 is the standard correction for continuity (Hays, 1973) and $c$ is the criterion. Next, obtain the probability, $\hat{p}_0$, that a standardized normal variate is less than $\hat{z}$, by using tables of normal distribution. Once this value is obtained, find the probability, $\hat{p}_{oo}$, that two standardized normal variates with correlation $\alpha_{21}$, are less than $\hat{z}$ by using tables of the bivariate normal distribution. Finally, substitute the obtained values of $\hat{p}_0$ and $\hat{p}_{oo}$ into the following equations to get a simple normal approximation of $k$, denoted $\hat{k}$:

$$
\hat{k} = \frac{\hat{p}_{oo} - \hat{p}_0^2}{\hat{p}_0 - \hat{p}_0^2} \tag{19}
$$

To calculate the value of $p$, using these same values of $\hat{p}_0$ and $\hat{p}_{oo}$, substitute them into the following equation
to yield a simple normal approximation of $p$, denoted $\hat{p}$:

$$\hat{p} = 1 + 2(\hat{p}_{ou} - \hat{p}_o)$$  \hspace{1cm} (20)

The results of this study by Peng and Subkoviak (1980) show that the simple normal approximation provides very accurate estimates of the exact values of $p$ and estimates with about 10% bias for $k$. There is support in the literature for the use of the normal distribution in estimating the beta-binomial family of distributions: (Shellman, 1948, p.260) shows that the negative binomial distribution can be used to approximate the beta-binomial distributions, and both Anscombe (1948) and Bartko (1966) demonstrate that the negative binomial distribution can be estimated by the normal family.

**Subkoviak's $\hat{p}$**

Subkoviak (1978a) proposes a single-test administration estimate of the probability of consistent mastery decisions if two tests had been administered. This is accomplished by substituting the assumptions of independent distribution of test scores and identical binomial distributions for fixed examinee $i$, as compared to Huynh's assumption of the bivariate distribution $X$ and $Y$, the group being beta-binomial and his estimate of a simple function $p$ known as the kappa coefficient (Subkoviak, 1984).

Subkoviak defines the coefficient of agreement for an individual $i$ as the probability that $i$ is assigned to the
same mastery state on parallel tests $X$ and $Y$. He states that the only two classifications possible in mastery-nonmastery tests $X$ and $Y$ with criterion $c$ are (1) $X \geq c$ and $Y \geq c$ indicating consistent mastery/mastery decisions and (2) $X < c$ and $Y < c$ indicating consistent nonmastery/nonmastery decisions. Subkoviak defines the coefficient of agreement for examinee $i$, when the criterion equals $c$ as:

$P_c = p[(X_i \geq c) (Y_i \geq c)] + p[(X_i < c) (Y_i < c)]$ \hspace{1cm} (21)

By using the assumption that test scores $X$ and $Y$ are independently distributed for a fixed person and that the distributions of $X$ and $Y$ for a fixed person are binomial in form, the following equation is derived:

$P_c = [p(X \geq c)] + [1 - p(X \geq c)]$ \hspace{1cm} (22)

where

$p(X_i \geq c) = \sum_{i=1}^{n} p_i X_i - X_i \sum_{i=1}^{n} p_i (1-p_i) n-X_i$ \hspace{1cm} (23)

where

$p_i = \alpha_{21}(X_i^n) + (1 - \alpha_{21}) \frac{M}{n_X}$ \hspace{1cm} (24)

The latter three equations can be used to determine the probability of consistent classification for each individual examinee.

The coefficient of agreement $P_c$ for a group of examinees of $N$ persons as:
Equation 25 is the sum of the probabilities of making a consistent decision for examinee \( i \) weighted by his probability of occurrence in the group, or by chance, and therefore reflects the probability of consistent decisions across the group and can be calculated from a single test administration.

Huynh (1977) later suggested a kappamax index of reliability of mastery-nonmastery classifications as having an advantage over the ordinary kappa in that the kappamax is not a function of the cutoff or criterion score, making it unique for a test given to a group of examinees. This index is not used in any of the literature, however.

These indices proposed by Hambleton and Novick (1973), Swaminathan et al (1974), Huynh (1976b), and Subkoviak (1978b) are all examples of the threshold loss function of reliability for criterion-referenced tests.

Previous Comparative Studies

Subkoviak (1978b) was the first to actually do a comparative study of various criterion-referenced reliability indices. He looked at the indices proposed by Swaminathan et al (1974), Huynh (1976a), Subkoviak (1978a) and the unpublished index of Marshall and Haertel.
Subkoviak used a data base of 1586 students' responses to parallel forms of 10, 30, and 50 items each from the Scholastic Aptitude Test (SAT). Items on the two 50-item forms were selectively deleted to create the two 30-item forms and items from the 30-item forms were similarly deleted to form the 10-item forms. The items for each form were taken from the Verbal section of the SAT, using approximately equal item difficulty and item discrimination across all forms. The means, standard deviations, and KR-21 reliabilities were then calculated for each form which resulted in distribution of scores for 1586 students on these parallel tests.

Four different mastery criteria or cut-off scores were considered for each n-item test: c- 50%, 60%, 70%, and 80% correct. The population parameter, the proportion ($P_c$) of the 1586 students consistently classified as master/master or nonmaster/nonmaster for parallel forms, were then calculated for each combination of test lengths, n and the four mastery criteria, c.

Subkoviak concluded that all four procedures "appear to provide reasonably accurate estimates of $P_c$, the proportion of consistent classifications on two mastery test, for the various cases considered" (Subkoviak, 1978b, p. 115). He did note the following advantages and disadvantages: The Swaminathan procedure produced unbiased estimates, but it required two testings and results in relatively large
standard errors for classroom size samples; the Huynh, Marshall-Haertel, and Subkoviak procedures required only one testing and produced relatively small standard errors, but for short tests, each procedure appeared to produce a unique kind of systematic bias. Subkoviak's final conclusion was that Huynh's approach appeared to be the best in that it is "Mathematically sound, required only one testing, and produced reasonably accurate estimates, which appeared to be slightly conservative for short tests" (Subkoviak, 1978, p. 115).

Marshall and Serlin (1979) conducted another comparative study, using the four single-administration indices of Brennan and Kane (1977a), Huynh (1976b), Marshall and Haertel, and Subkoviak (1978a) as the basis of their investigation. The procedure used was to generate item-by-examinee response matrices, according to various parameters selected to control for score distribution shape, modality, mean, and variance. A score distribution was obtained for each matrix and test indices were then calculated for each integral cut-off score. The authors of this study considered five types of score distributions: bell-shaped, highly negatively skewed, non-symmetrical bimodal, symmetrical bimodal with modes well separated, and symmetrical bimodal with modes near each other.

Marshall and Serlin concluded that the index of dependability developed by Brennan and Kane was clearly
different from the other indices studied and it clearly measured different things. Coefficient kappa did not reflect score distribution modes whereas Huynh's \( p \) did generally reflect unimodal score distributions very well except with modes close together. The investigators stated that Subkoviak's procedure, then, should not be used for bimodal distributions; instead, it should only be used for unimodal distributions that closely approximate a normal distribution. Marshall and Serlin concluded that the question should not be whether to use the Huynh \( p \) or Subkoviak \( P_c \) or Marshall-Haertel \( \beta \), but instead the question should be which set of assumptions is fitting for the situation (Marshall and Serlin, 1979).

Huynh and Saunders (1980a) compared the beta-binomial estimates of the raw agreement index \( p \) and corrected-for-chance kappa index, \( k \), as calculated by Huynh's single-test administration with estimates based on two test administrations.

Comparisons were made with \( n = 5, 10, 20, \) and 30 test items. A test mean and KR-21 reliability were chosen that would result in test score distributions of U-shaped, symmetric, unimodal, or J-shaped. The criterion, \( c \), was chosen such that \( c/n \) would be 60\%, 70\%, or 80\%. Data was collected from a portion of the Comprehensive Tests of Basic Skills (CTBS) by taking every tenth case from the entire
South Caroline file. This resulted in a number of examinees ranging from \( N = 1684 \) to 6035.

The authors concluded that the single administration (beta-binomial) estimate for the raw agreement index \( p \) provided a negligible amount of negative bias and the kappa index displayed a moderate degree (about ten percent) of negative bias. The investigators did state, however, that the "small bias associated with the beta-binomial estimates for \( p \) and \( k \) does not imply that decision reliability should always be based on one test administration. In most reliability studies, estimates based on two test administrations are preferable... However, when two test administrations are not feasible, then single administration estimated based on the beta-binomial may be used" (Huynh and Saunders, 1980a, p. 357).

Use Of An Indifference Zone

An indifference zone is defined by Wilcox (1976) as an area between two cut-off scores where a test user can be "indifferent" as to how examinees are classified. He assumes that there will be negligible loss of misclassifying such examinees, but one must be reasonably certain that examinees with scores greater than the upper cut-off score or lower than the lower cut-off score are correctly classified. Huynh (1980b and 1980c) and Wilcox (1979) both provide means of calculating the probability of making false positive and false negative errors using and indifference zone, but no
studies have been conducted to investigate which reliability index is most appropriate for these studies.

Loss Due To Misclassifications

The squared-error loss approach to measurement error does not assume the losses associated with false mastery and false nonmastery error are equally serious. It considers the consequence of misclassifying examinees far above or far below the cut-off score to be more serious than misclassifying those at or near the cut-off score (Berk, 1980, chapter 9).

In addition to the use of the squared-error loss function, Cohen (1968) introduced a weighted kappa which is the "proportion of weighted agreement corrected for chance, to be used when different kinds of disagreement are to be differentially weighted in the agreement index" (p. 215). Cohen computed $k$ and $k_m$ on a group of 200 examinees where all disagreements are given the same weight. In other words, $k_m$ is a generalization of $k$, proportion of weighted agreement corrected for chance.

Although this type of function (squared-error loss) was the first reliability index to be considered, it is used the least in current literature.

Degree of Mastery or Nonmastery

If different degrees of mastery and nonmastery exist to any great extent, the threshold loss function may not be
appropriate since it is mainly a qualitative distinction. Once an initial classification as master or nonmaster has been made based on a cut-off score, the information of the greatest interest to teachers is frequently the degree of mastery or nonmastery of the student along the score continuum (Berk, 1984, chapter 9). Under these conditions, then, another loss function must be employed (Brennan and Kane, 1980). The general index of choice has been the squared-error loss agreement indices $k^2(X, T_x)$ or $M(C)$ (van der Linden and Mellenberg, 1978).
CHAPTER BIBLIOGRAPHY


CHAPTER 3

METHOD OF RESEARCH

The procedure used in this study was one that was suggested by Marshall and Serlin (1979): computer-generated item-by-examinee response (correct or incorrect) matrices were obtained, according to various parameters chosen to control score distribution shape, modality, mean, and variance. From each matrix, test indices were calculated for all chosen cut-off scores or criteria.

This study combined all the areas examined in the three previous published comparative studies of Subkoviak (1978), Marshall and Serlin (1979), and Huynh and Saunders (1980a), in addition to other areas, to give a thorough comparison of the relative merits of Livingston's $k^2(X,T)$, Brennan and Kane's $M(C)$, Subkoviak's $P_e$, Huynh's $p$ and $k$, and Swaminathan's $p$ and $k$.

Description of Simulation

The first step in the generation of random data was to randomly generate values of 0 to 10, recoded later, for $n = 60$ items and $N = 2000$ cases for both normal and negatively-skewed distributions. The second step was to generate a normal distribution of the $N$ cases between the
values of 1 to 10 as well as a negatively-skewed distribution for the same N cases and the same values of 1 to 10. These values represent an examinee's true probability of success in correctly answering items, with 0 representing the probability of getting 0 out of 10 items correct; 1 being the probability of 1 out of 10 answered correctly; 2 representing 2 out of 10 correct; ... 10 representing the probability of 10 out of 10 answered correctly.

Once the normal distribution of values was obtained, the original random data (N = 2000, n = 60) was recoded such that the correct number of cases (as determined earlier) with the true probability of 0 had scores with all digits recoded to 0 to represent "incorrect"; the correct number of cases with the true probability of 1 had digits 2 through 10 recoded to 0 (incorrect) and digits of 1 were recoded to 1 (correct); the correct number of cases with the true probability of 2 had digits 3 through 10 recoded to 0 (incorrect) and the digits 1 through 2 were recoded to 1 (correct). This process was continued until all 2000 cases were recoded in a normally distributed manner.

This entire recoding process was repeated in the same manner for the negatively-skewed distribution's true probability of success as determined by the randomly generated distribution data described earlier.

The scores were then summed for this group of 2000 examinees on each of 2 randomly parallel tests for each
combination of test length (n) and master criteria (c).
These were summed as follows: Items 1-5 and items 6-10
represent scores on two n = 5 tests; items 1-10 and 11-20
represent scores on two n = 10 tests; and items 1-20 and
21-40 represent scores on two n = 20 tests. For each
combination of test length (n) and criterion (c), the
proportion, P, of examinees classified consistently as master
or nonmaster on both test forms was calculated. Both the
normally distributed and negatively-skewed distribution
scores were saved in separate 6 x 2000 matrices (6 tests x
2000 examinees) for further analysis.

Once these population values for the proportion, P, of
consistent master/nonmaster classifications were calculated,
samples were drawn from the population of 2000 for each of
the six combinations of examinees (N = 30, 150) and number of
items (n = 5, 10, 20). For each of the two different cutoff
scores or master criteria, c, (60%, 80%), the different
reliability indices were then calculated for each
combination, using each of the seven methods being compared.
This method was repeated 1000 times from which a mean index
was calculated, as well as the descriptive statistics of each
reliability index.

The following formulas were used:

   a. P
\( p = \sum_{i=1}^{k} P_{ii} \), where \( P_{ii} \) is the proportion of examinees classified in the \( i \)th mastery level on both tests.

\[ k = \frac{(P_o - P_c)}{1 - P_o} \]

where \( P_o = \sum_{i=1}^{k} P \) and \( P_c = \sum_{i=1}^{k} P_i \). \( P_i \), where \( P_i \) is the proportions of examinees assigned to mastery state \( i \) on the first and second test administrations.

2. Huynh (1976b)

a. kappa (normal approximation suggested by Peng and Subkoviak, 1980)

\[ z = \frac{(c - .5 - \mu)}{\sigma} \]

where .5 is the standard correction for continuity (Hays, 1973, p.309).

Next, using tables of normal distribution, obtain the probability, \( \hat{P}_o \), that a standardized normal variate is less than \( \hat{z} \). Next, using tables of the bivariate normal distribution, obtain the probability, \( \hat{P}_{oo} \), that two standardized normal variates with reliability \( \alpha_{21} \) are less than \( \hat{z} \) in the above equation. Finally, substitute the values \( \hat{P}_o \) and \( \hat{P}_{oo} \) into the following equation to yield simple normal approximations of \( k \), denoted
\[ \hat{k} = \frac{(\hat{p}_{oo} - \hat{p}_o^2)}{(\hat{p}_o - \hat{p}_o^2)} \]

b. \( p \) (normal approximation)

Substitute the same values of \( \hat{p}_{oo} \) and \( \hat{p}_o \) found in (a), into the following equation to yield simple normal approximations of \( p \), denoted \( \hat{p} \):

\[ \hat{p} = 1 + 2(\hat{p}_{oo} - \hat{p}_o) \]

3. Subkoviak (1978a) \[ \frac{P_c}{N} \]

\[ P_c = \sum_{i=1}^{N} P_{c(i)} \]

where \( P_{c(i)} = [P(X_i \geq C)]^2 + [1 - P(X_i \geq C)]^2 \)

where \( P(X_i \geq C) = \sum_{X=C}^{n} \binom{n}{X} p_i X_i (1 - p_i)^{n-X_i} \)

where \( p_i = \alpha_{Z1} \frac{X_i}{n} + (1 - \alpha_{Z1}) \frac{P_c}{C_x} \)

4. Livingston (1972) \[ k^2(x, \overline{I}_x) \]

\[ k^2(x, \overline{I}_x) = \alpha_{Z1} \frac{(\sigma^2)}{\sigma^2 + (\mu_x - C_x)^2} \]

5. Brennan and Kane (1977a) \[ M(C) \]

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

Once these calculations are made, the mean and standard error of each index was calculated and compared to the
population value for that combination. The deviation of the estimated value from the parameter value was then determined.

To answer research question 4, the indices were examined to determine which one exhibited the highest reliability when two separate cutoff scores were considered for one test; in other words, which one had the combined highest reliability when the criterion of 60% and the criterion of 80% were considered together.

To answer research question 5, the indices of Livingston and Brennan and Kane were examined to determine if the squared-error loss function did indeed have the advantage of being sensitive to the magnitude of errors, whereas the threshold loss function did not.
CHAPTER BIBLIOGRAPHY


CHAPTER 4

PRESENTATION AND ANALYSIS OF DATA

This study was conducted in order to determine the accuracy and consistency of decisions made from criterion-referenced tests in classifying examinees as either masters or nonmasters. This was done by examination of the sampling distribution characteristics and comparing seven methods of computing reliability for criterion-referenced tests. The seven methods computed were (1) Livingston's $k^2(X,T_X)$, (2) Brennan and Kane's $M(C)$, (3) Huynh's $p$, (4) Huynh's $k$, (5) Subkoviak's $p$, (6) Swaminathan's $p$, and (7) Swaminathan's $k$. Computer-generated data produced scores for three parallel-forms of a test ($n=5, 10,$ and $20$ items) for two populations: A normally distributed ($N=2000$) and a negatively-skewed one ($N=2000$). These two test forms were generated such that the mean, variance, and correlations were comparable to actual parallel tests as shown by the data in Table 1. Having satisfied these requirements, $1000$ samples each of size $N=30$ and $N=150$ were drawn, with replacement, from the population of $2000$ cases and the seven methods of computing reliability were then calculated on each sample. This process was done from both the normally distributed population and the negatively-skewed population. The results of these calculations are reported in the
analysis section below.

Table 1

Descriptive Statistics for Two Populations (N=2000)

<table>
<thead>
<tr>
<th>Test Form</th>
<th>n</th>
<th>μ</th>
<th>σ²</th>
<th>r</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>2.9</td>
<td>1.71</td>
<td>.375</td>
<td>.78</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2.9</td>
<td>1.73</td>
<td>.375</td>
<td>.74</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>5.8</td>
<td>4.71</td>
<td>.537</td>
<td>.598</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>5.9</td>
<td>4.32</td>
<td>.537</td>
<td>.598</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>11.7</td>
<td>13.88</td>
<td>.697</td>
<td>.738</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>11.8</td>
<td>14.44</td>
<td>.697</td>
<td>.738</td>
</tr>
</tbody>
</table>

Note. μ = mean and σ² = variance
r = correlation between Form 1 & 2

Analysis

Research Question 1a

What are the sampling distribution characteristics of Livingston's \( k^2(X, T_x) \), a squared-error loss function, as used in determining the reliability of criterion-referenced tests?

The descriptive statistics of Livingston's \( k^2(X, T_x) \) for the 1000 samples of N=30 and N=150 from the normal population are shown in Table 2. Examination of this table reveals
that when holding the number of examinees (N) and the

Table 2

Descriptive Statistics for Livingston's $k^2(x, T)$: Normal Population

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>

Note. $Q_1$ = skewness and $Q_2$ = kurtosis

criterion ($c$) constant, with increasing items (n) the value of Livingston's index increases and the standard deviation and range generally decrease. The large range and
standard deviation for $n=5$, $N=30$, and $c=60\%$ are due to the fact that the reliability, under these conditions, had a negative value on several occasions. According to Livingston, this can occur when the cutoff score and the mean are close together (1972).

The increases in the values of the index are more dramatic from $n=5$ to $n=10$ than from $n=10$ to $n=20$ due mainly to the larger difference between the criterion and the population mean due to more items in the test. Under these same conditions, the skewness value moves toward 0 and the kurtosis value increases, a move toward more normal curve characteristics. The exception to this is at $n=20$ where the reverse is true.

Comparing the two sample sizes $N=30$ and $N=150$, while holding $c$ and $n$ constant, the same general trends are observed: the $k^2(X, I_x)$ value generally increases with the standard deviation and the range decreasing as the sample size ($N$) increases. As $N$ increases, the mean is closer to the median point as well.

With an increase in sample size ($N$), the index generally increases with a decrease in the standard deviation and the range, because of the increasing distance between the population mean and the criterion. The skewness generally shows a move to a more negative value under these conditions, however, as the kurtosis moves toward a more positive value, except when $n=20$ and $N=150$ when there is a slight decrease.
At c=80%, n=20, and N=150, the range is very small, with the first and second quartile having the same value.

Examination of Table 3 reveals the same trends.

Table 3

**Descriptive Statistics for Livingston's $k^2(X, T_x)$: Negatively-skewed Population**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>$n$</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>60%</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td>80%</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>

Note: $q_1 = \text{skewness}$ and $q_2 = \text{kurtosis}$
for the samples from the negatively-skewed population as were evident in the samples from the normally-distributed population when \( N \) and \( n \) are varied separately. When \( n=20 \), \( N=150 \), and \( \sigma=60\% \) there is a very small range and each quartile shows the same value. When the criterion is varied with \( N \) and \( n \) held constant, the reverse of the normally-distributed population is true: the value of \( k^2(X,\bar{X}) \) decreases with the standard deviation and range increasing. This is due to the presence of the \((\mu_x - \bar{C}_x)^2\) present in the formula. In a negatively-skewed population, the mean will be closer to the value of the cutoff score, causing \( k^2(X,\bar{X}) \) to be lower. The distribution of the population from which samples are drawn does not affect the calculation of Livingston's reliability index until the criterion is increased, in which case there will be a difference between the normal and negatively-skewed populations. In the negatively-skewed population the mean is also much closer to the median than in the normal population.

Research Question 1b

What are the sampling distribution characteristics of Brennan and Kane's \( M(C) \), a squared-error loss function, as used in determining the reliability of criterion-referenced tests?

The descriptive characteristics of \( M(C) \) are listed in Table 4 for the normal population and Table 5 for the negatively-skewed population. Close examination of these
Tables reveal that the statistics are identical to Livingston's statistics. Even though a very different formula was used to calculate each index of reliability, they are algebraically equivalent (Appendix G). Because of Table 4

Descriptive Statistics for Brennan and Kane's M(C): Normal Population

<table>
<thead>
<tr>
<th>g</th>
<th>n</th>
<th>N</th>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>60%</td>
<td>3 5</td>
<td>150</td>
<td>$\mu$</td>
<td>0.362</td>
</tr>
<tr>
<td>6 10</td>
<td>30</td>
<td></td>
<td>$\sigma$</td>
<td>0.313</td>
</tr>
<tr>
<td>6 10</td>
<td>150</td>
<td></td>
<td>range</td>
<td>0.32</td>
</tr>
<tr>
<td>12 20</td>
<td>30</td>
<td></td>
<td>$g_1$</td>
<td>-1.159</td>
</tr>
<tr>
<td>12 20</td>
<td>150</td>
<td></td>
<td>$g_2$</td>
<td>0.23</td>
</tr>
<tr>
<td>80%</td>
<td>4 5</td>
<td>30</td>
<td>$\mu$</td>
<td>0.610</td>
</tr>
<tr>
<td>4 5</td>
<td>150</td>
<td></td>
<td>$\sigma$</td>
<td>0.618</td>
</tr>
<tr>
<td>8 10</td>
<td>30</td>
<td></td>
<td>$g_1$</td>
<td>0.763</td>
</tr>
<tr>
<td>8 10</td>
<td>150</td>
<td></td>
<td>$g_2$</td>
<td>0.769</td>
</tr>
<tr>
<td>16 20</td>
<td>30</td>
<td></td>
<td>$\mu$</td>
<td>0.864</td>
</tr>
<tr>
<td>16 20</td>
<td>150</td>
<td></td>
<td>$\sigma$</td>
<td>0.863</td>
</tr>
</tbody>
</table>

Note: $g_1$ = skewness and $g_2$ = kurtosis
these identical values, the same trends are true for $M(C)$ as are true for $k^{2}(X, x)$. For example, the value of the reliability index increases with increasing number of items,

Table 5

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>3 5 30</td>
<td>0.846</td>
</tr>
<tr>
<td>3 5 150</td>
<td>0.848</td>
</tr>
<tr>
<td>6 10 30</td>
<td>0.917</td>
</tr>
<tr>
<td>6 10 150</td>
<td>0.920</td>
</tr>
<tr>
<td>12 20 30</td>
<td>0.958</td>
</tr>
<tr>
<td>12 20 150</td>
<td>0.959</td>
</tr>
<tr>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>4 5 30</td>
<td>0.540</td>
</tr>
<tr>
<td>4 5 150</td>
<td>0.549</td>
</tr>
<tr>
<td>8 10 30</td>
<td>0.722</td>
</tr>
<tr>
<td>8 10 150</td>
<td>0.719</td>
</tr>
<tr>
<td>16 20 30</td>
<td>0.836</td>
</tr>
<tr>
<td>16 20 150</td>
<td>0.836</td>
</tr>
</tbody>
</table>

Note. $q_{1}$ = skewness and $q_{2}$ = kurtosis
with increasing sample size, and with increasing criterion.

Research Question 2a

What are the sampling distribution characteristics of Huynh's $p$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

The 1000 samples of both the normally distributed population and the negatively-skewed one are arranged in Tables 6 and 7 respectively. Upon examination of the values in Table 6, one notes that when the sample size ($N$) increases, the value of $p$ decreases very slightly in all cases but one ($n=20, c=60\%$), and yet the standard deviation decreases, as does the range. Under these conditions, the mean closely approximates the median. The skewness approaches normality with increased $N$ and the kurtosis dramatically decreases. The range decreases considerably, also, with increased $N$, while the variation in the quartiles is greater with decreased $N$. When the number of items ($n$) increases, however, and the other variables are held constant, the value of $p$ does increase while the standard deviation and the range decrease, as expected. The skewness of the distribution, when the criterion is 60%, varies irregularly; but, when the criterion is 80% it becomes less negative, moving toward a normal distribution. This same pattern is followed with the kurtosis, in that it becomes generally less positive. When the criterion increases from 60% to 80% the value of $p$ also increases, with decreasing standard deviation
and range noted. For a small \( n \), the skewness becomes more negative with increasing criterion, but for a larger \( n \)

Table 6

Descriptive Statistics for Huynh's \( p \):

| Normal Population |

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>( \sigma )</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>1 6</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>90%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>

Note. \( q_1 \) = skewness and \( q_2 \) = kurtosis

(10 or 20), it becomes less negative. The kurtosis value
shows a steady increase as the criterion increases. Under these conditions of increased criterion, the median value goes from slightly below the mean at 60% to slightly above the mean at 80%. The difference in the quartile values is generally less pronounced with the higher criterion.

For the negatively-skewed population, analysis of Table 7 shows that the characteristics of the values of $p$ are slightly different than for the normally distributed population. With an increase in sample size ($N$), these values do increase as expected when the lower criterion is examined. When the criterion is higher (80%), the values decrease with increased sample size. In both instances, the standard deviation and the range decreases, with the skewness and kurtosis moving more toward a normal distribution curve. The mean more closely approximates the median at increased $N$ also. When the number of items is increased with the other variables constant, there is also an increase in the value of Huynh's $p$ index, with a very slight decrease in the standard deviation and the range of values. The skewness and kurtosis values indicate a move away from the normal curve to a more negative value for the skewness at $c=60\%$ and a more positive value at $c=80\%$ and a generally more positive value for the kurtosis.

When only the criterion is considered, the values of $p$ decrease as the criterion increases, with a slight increase in the standard deviation and the range. The
skewness continually becomes less negative and the kurtosis more positive. The same trend as with the normal data follows here with the mean being below the median at the 60% criterion and above the median at a criterion of 80%.

Table 7

Descriptive Statistics for Huynh's p:
Negatively-skewed Population

<table>
<thead>
<tr>
<th>c</th>
<th>n</th>
<th>N</th>
<th>µ</th>
<th>σ</th>
<th>range</th>
<th>Q₁</th>
<th>Q₂</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>3</td>
<td>5</td>
<td>30</td>
<td>.971</td>
<td>.029</td>
<td>.17</td>
<td>-1.292</td>
<td>1.723</td>
<td>.96</td>
<td>.98</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>150</td>
<td>.975</td>
<td>.013</td>
<td>.07</td>
<td>-.616</td>
<td>-.010</td>
<td>.97</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>30</td>
<td>.976</td>
<td>.023</td>
<td>.13</td>
<td>-1.318</td>
<td>1.789</td>
<td>.96</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>150</td>
<td>.981</td>
<td>.010</td>
<td>.06</td>
<td>-.636</td>
<td>.461</td>
<td>.98</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>30</td>
<td>.987</td>
<td>.015</td>
<td>.08</td>
<td>-1.461</td>
<td>2.325</td>
<td>.98</td>
<td>.99</td>
<td>1.0</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>150</td>
<td>.989</td>
<td>.006</td>
<td>.03</td>
<td>-.336</td>
<td>.526</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>80%</td>
<td>4</td>
<td>5</td>
<td>30</td>
<td>.827</td>
<td>.074</td>
<td>.38</td>
<td>.212</td>
<td>-.213</td>
<td>.78</td>
<td>.83</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>150</td>
<td>.816</td>
<td>.030</td>
<td>.21</td>
<td>.236</td>
<td>.335</td>
<td>.80</td>
<td>.81</td>
<td>.84</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>30</td>
<td>.831</td>
<td>.052</td>
<td>.33</td>
<td>.483</td>
<td>.327</td>
<td>.79</td>
<td>.83</td>
<td>.86</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>150</td>
<td>.821</td>
<td>.021</td>
<td>.15</td>
<td>.275</td>
<td>.168</td>
<td>.81</td>
<td>.82</td>
<td>.84</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>30</td>
<td>.858</td>
<td>.040</td>
<td>.26</td>
<td>.468</td>
<td>.341</td>
<td>.83</td>
<td>.85</td>
<td>.88</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>150</td>
<td>.852</td>
<td>.016</td>
<td>.24</td>
<td>.280</td>
<td>.152</td>
<td>.84</td>
<td>.85</td>
<td>.86</td>
</tr>
</tbody>
</table>

Note: $q_1 = \text{skewness}$ and $q_2 = \text{kurtosis}$
Research Question 2b

What are the sampling distribution characteristics of Huynh's k, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

The descriptive statistics for the 1000 samples were calculated using the procedure for Huynh's k and are shown in Table 8 for the normally distributed population and in Table 9 for the negatively-skewed population. The first glance immediately makes it obvious that these are very low numbers for a reliability index. Further examination reveals that the mean value of k increases with increasing sample size, while the standard deviation and range of scores decreases, a trend one can usually expect with increased N. The mean also falls closer to the median with increased sample size as well. If the number of items (n) is increased, the value of k also increases with the standard deviation and range of scores decreasing at a steady rate. The difference between the first and third quartiles decreases with increasing n also. With an increase from n=5 to n=10 there is a trend toward more normal curves as evidenced by the values for the skewness and kurtosis, but there is a move away from the normal curve proportions when increasing from n=10 to n=20.

When considering the criterion (c), as it increases, the value of Huynh's kappa decreases very slightly, with the standard deviation and the range either holding constant or
Table 8

**Descriptive Statistics for Huyoh's k:**
*Normal Population*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>n</td>
</tr>
<tr>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>

Note. $g_1$ = skewness and $g_2$ = kurtosis.

Increasing slightly. The same trends that were evident in the distribution of the k's with increasing number of items are also evident with increasing criterion.
Analysis of Table 9 shows that the values of the descriptive statistics for the 1000 samples of the negatively-skewed population are generally even lower than Table 9

**Descriptive Statistics for Huynh's k: Negatively-skewed Population**

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>μ</th>
<th>σ</th>
<th>Range</th>
<th>q₁</th>
<th>q₂</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 5 30</td>
<td>.123</td>
<td>.187</td>
<td>3.25</td>
<td>10.378</td>
<td>161.491</td>
<td>.02</td>
<td>.09</td>
<td>.19</td>
</tr>
<tr>
<td>3 5 150</td>
<td>.107</td>
<td>.053</td>
<td>.27</td>
<td>.429</td>
<td>-.377</td>
<td>.06</td>
<td>.10</td>
<td>.14</td>
</tr>
<tr>
<td>6 10 30</td>
<td>.198</td>
<td>.126</td>
<td>.54</td>
<td>.352</td>
<td>-.713</td>
<td>.10</td>
<td>.19</td>
<td>.30</td>
</tr>
<tr>
<td>6 10 150</td>
<td>.200</td>
<td>.057</td>
<td>.34</td>
<td>-.019</td>
<td>-.161</td>
<td>.16</td>
<td>.20</td>
<td>.24</td>
</tr>
<tr>
<td>12 20 30</td>
<td>.280</td>
<td>.128</td>
<td>.58</td>
<td>-.047</td>
<td>-.746</td>
<td>.18</td>
<td>.28</td>
<td>.38</td>
</tr>
<tr>
<td>12 20 150</td>
<td>.294</td>
<td>.057</td>
<td>.34</td>
<td>-.257</td>
<td>-.107</td>
<td>.26</td>
<td>.30</td>
<td>.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>μ</th>
<th>σ</th>
<th>Range</th>
<th>q₁</th>
<th>q₂</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 5 30</td>
<td>.202</td>
<td>.138</td>
<td>.50</td>
<td>.254</td>
<td>-.744</td>
<td>.09</td>
<td>.20</td>
<td>.30</td>
</tr>
<tr>
<td>4 5 150</td>
<td>.225</td>
<td>.062</td>
<td>.36</td>
<td>-.044</td>
<td>-.246</td>
<td>.18</td>
<td>.22</td>
<td>.27</td>
</tr>
<tr>
<td>8 10 30</td>
<td>.357</td>
<td>.116</td>
<td>.70</td>
<td>-.489</td>
<td>.158</td>
<td>.28</td>
<td>.37</td>
<td>.44</td>
</tr>
<tr>
<td>8 10 150</td>
<td>.375</td>
<td>.048</td>
<td>.26</td>
<td>-.283</td>
<td>-.170</td>
<td>.34</td>
<td>.38</td>
<td>.41</td>
</tr>
<tr>
<td>16 20 30</td>
<td>.484</td>
<td>.090</td>
<td>.61</td>
<td>-.865</td>
<td>1.011</td>
<td>.43</td>
<td>.50</td>
<td>.55</td>
</tr>
<tr>
<td>16 20 150</td>
<td>.501</td>
<td>.033</td>
<td>.21</td>
<td>-.346</td>
<td>.264</td>
<td>.48</td>
<td>.50</td>
<td>.52</td>
</tr>
</tbody>
</table>

*Note*: $q_1$ = skewness and $q_2$ = kurtosis.
the values for the normal population. The values increase with increasing N in all but one instance, when n=5 and c=60%. The standard deviation and the range of values decreases every time N is increased. Further examination of the raw data for n=5, N=30, and c=60% revealed two outliers that were obviously responsible for the unusual values at this point. The mean value is very close to that of the median at most points, with the difference between the first and third quartiles relatively steady. The skewness values vary from situation to situation, with no real pattern easily discerned. The kurtosis is rather erratic from situation to situation also, all of which seem to support the literature assumptions that Huynh's k is not a very good estimator for negatively-skewed populations (Marshall and Serlin, 1979).

Further analysis shows that both an increase in n and c respectively, result in increases in the value of k, slight decreases in the standard deviation and range and a move away from normal curve characteristics with more negative skewness values and more positive kurtosis values in nearly every instance of increased n and increased c.

Research Question 2c

What are the sampling distribution characteristics of Subkoviak's $P_n$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

The characteristics of Subkoviak's $P_n$ are shown in Tables 10 and 11, the former showing those for the normal
distribution and the latter for the negatively-skewed
distribution. As sample size increases, as shown in Table
10, the value of $P_c$ increases slightly except when $c=80$
and

Table 10

**Descriptive Statistics for Subkoviak's $P_c$: Normal Population**

<table>
<thead>
<tr>
<th>$c$</th>
<th>$n$</th>
<th>$N$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>range</th>
<th>$q_1$</th>
<th>$q_2$</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>3</td>
<td>5</td>
<td>30</td>
<td>0.546</td>
<td>0.064</td>
<td>0.34</td>
<td>-0.411</td>
<td>1.204</td>
<td>0.51</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>150</td>
<td>0.557</td>
<td>0.025</td>
<td>0.15</td>
<td>-0.025</td>
<td>-0.021</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>10</td>
<td>30</td>
<td>0.556</td>
<td>0.076</td>
<td>0.46</td>
<td>0.113</td>
<td>-0.248</td>
<td>0.50</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>10</td>
<td>150</td>
<td>0.560</td>
<td>0.032</td>
<td>0.19</td>
<td>0.050</td>
<td>-0.156</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>20</td>
<td>30</td>
<td>0.530</td>
<td>0.082</td>
<td>0.28</td>
<td>-0.016</td>
<td>0.245</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>20</td>
<td>150</td>
<td>0.535</td>
<td>0.034</td>
<td>0.20</td>
<td>0.119</td>
<td>-0.131</td>
<td>0.51</td>
<td>0.53</td>
</tr>
<tr>
<td>80%</td>
<td>4</td>
<td>5</td>
<td>30</td>
<td>0.242</td>
<td>0.043</td>
<td>0.26</td>
<td>-0.003</td>
<td>0.024</td>
<td>0.21</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>150</td>
<td>0.244</td>
<td>0.019</td>
<td>0.14</td>
<td>0.048</td>
<td>0.340</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>10</td>
<td>30</td>
<td>0.189</td>
<td>0.056</td>
<td>0.36</td>
<td>0.187</td>
<td>-0.036</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>10</td>
<td>150</td>
<td>0.186</td>
<td>0.025</td>
<td>0.16</td>
<td>-0.021</td>
<td>-0.033</td>
<td>0.17</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>20</td>
<td>30</td>
<td>0.133</td>
<td>0.055</td>
<td>0.30</td>
<td>0.456</td>
<td>-0.134</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>20</td>
<td>150</td>
<td>0.132</td>
<td>0.023</td>
<td>0.15</td>
<td>0.088</td>
<td>-0.045</td>
<td>0.12</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**Note.** $q_1$ = skewness and $q_2$ = kurtosis
n=10 and n=20, in which case there is a very slight decrease. The standard deviation and range of values decreases steadily. For the lower criterion, the skewness becomes a less negative value from n=5 and n=10, but increases when n=20. When c=80% the skewness approaches normality for n=10 and n=20 but moves further away when n=5. The kurtosis shows a more positive value under the same conditions, which is a move away from normality. The range between quartiles decreases with a greater N.

With an increase in n, while holding other variables constant, the value of $P_c$ decreases with every increase except from n=5 to n=10 with c=60%, with the standard deviation and range of scores generally increasing. When examining the skewness and kurtosis columns of Table 10, it is apparent that Subkoviak's values exhibit a great degree of the normal curve characteristics, with low negative values for skewness and low positive values for kurtosis. As the criterion increases, however, there is a drastic decrease in the value of $P_c$. The mean falls relatively close to the median at all values except at n=20, where it is greater than the median.

The values of $P_c$ are displayed in Table 11 for the 1000 samples drawn from the negatively-skewed population. Examination reveals that these values follow the same general trends that those from the normally distributed population exhibit. With a larger sample size (N) the mean
falls closer to, and sometimes is greater than, the median, although the median value remains constant in most cases, while the value of $P_c$ increases slightly with sample size, the standard deviation decreases by nearly half. The range

Table 11

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>range</th>
<th>$q_1$</th>
<th>$q_2$</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>60%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>30</td>
<td>.372</td>
<td>.082</td>
<td>.52</td>
<td>- .162</td>
<td>.063</td>
<td>.31</td>
<td>.38</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>150</td>
<td>.382</td>
<td>.036</td>
<td>.22</td>
<td>- .049</td>
<td>-.031</td>
<td>.36</td>
<td>.38</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>30</td>
<td>.542</td>
<td>.091</td>
<td>.53</td>
<td>- .061</td>
<td>-.165</td>
<td>.48</td>
<td>.54</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>150</td>
<td>.545</td>
<td>.155</td>
<td>.25</td>
<td>- .122</td>
<td>-.130</td>
<td>.52</td>
<td>.55</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>30</td>
<td>.676</td>
<td>.092</td>
<td>.55</td>
<td>- .113</td>
<td>-.502</td>
<td>.61</td>
<td>.70</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>150</td>
<td>.679</td>
<td>.038</td>
<td>.24</td>
<td>- .176</td>
<td>.140</td>
<td>.65</td>
<td>.68</td>
</tr>
<tr>
<td><strong>80%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>30</td>
<td>.286</td>
<td>.058</td>
<td>.31</td>
<td>- .345</td>
<td>-.218</td>
<td>.25</td>
<td>.29</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>150</td>
<td>.294</td>
<td>.025</td>
<td>.15</td>
<td>- .188</td>
<td>-.020</td>
<td>.28</td>
<td>.29</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>30</td>
<td>.426</td>
<td>.070</td>
<td>.43</td>
<td>- .202</td>
<td>-.278</td>
<td>.37</td>
<td>.43</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>150</td>
<td>.432</td>
<td>.033</td>
<td>.21</td>
<td>- .200</td>
<td>.123</td>
<td>.41</td>
<td>.43</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>30</td>
<td>.533</td>
<td>.080</td>
<td>.58</td>
<td>- .121</td>
<td>.274</td>
<td>.48</td>
<td>.54</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>150</td>
<td>.538</td>
<td>.032</td>
<td>.22</td>
<td>- .117</td>
<td>-.110</td>
<td>.52</td>
<td>.54</td>
</tr>
</tbody>
</table>

*Note. $q_1$ = skewness and $q_2$ = kurtosis*
also decreases regularly. When \( n \) is increased there is a
fairly marked increase in the index value. The skewness and
kurtosis values continue to change in a rather sporadic
manner, although their values fairly closely approximate
normal values. There also is a fairly significant decrease
in the value of \( P^c \) when the criterion goes from 60% to 80%,
as well as in the value of the standard deviation and in the
range of values. The differences in the quartiles is more
marked with lower \( n \) and lower \( N \), but the mean value is very
close to the median value at all times.

**Research Question 2d**

What are the sampling distribution characteristics of
Swaminathan's \( P \), a threshold loss function, as used in
determining the reliability of criterion-referenced tests?

Examination of Tables 12 and 13 which display the values
for the 1000 samples from the normal and negatively-skewed
populations respectively, shows that these values of the
reliability index, \( P \), as calculated by Swaminathan, are
slightly larger than the previous indices from the respective
populations. The value of the index increases in a
consistent manner with increase in sample size, increase in
number of items, and with increase in the criterion. This
trend was expected since, as the literature points out, it is
the only one that uses a two-test procedure (Subkoviak,
1978b). The standard deviation also shows a steady decrease
with each of the previously mentioned decreases, as does the range of values. The skewness and kurtosis display patterns of approaching more normal characteristics with increasing

Table 12

Descriptive Statistics for Swaminathan’s p:
Normal Population

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25 50 75</td>
</tr>
<tr>
<td>c n N</td>
<td>μ  σ range</td>
</tr>
<tr>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>3 5 30</td>
<td>.647 .061 .50</td>
</tr>
<tr>
<td>3 5 150</td>
<td>.646 .059 .38</td>
</tr>
<tr>
<td>6 10 30</td>
<td>.700 .062 .52</td>
</tr>
<tr>
<td>6 10 150</td>
<td>.701 .061 .51</td>
</tr>
<tr>
<td>12 20 30</td>
<td>.739 .037 .21</td>
</tr>
<tr>
<td>12 20 150</td>
<td>.739 .040 .28</td>
</tr>
<tr>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>4 5 30</td>
<td>.642 .063 .44</td>
</tr>
<tr>
<td>4 5 150</td>
<td>.653 .043 .28</td>
</tr>
<tr>
<td>8 10 30</td>
<td>.748 .024 .18</td>
</tr>
<tr>
<td>8 10 150</td>
<td>.747 .048 .30</td>
</tr>
<tr>
<td>16 20 30</td>
<td>.849 .039 .28</td>
</tr>
<tr>
<td>16 20 150</td>
<td>.855 .028 .22</td>
</tr>
</tbody>
</table>

Note. θ₁ = skewness and θ₂ = kurtosis
values of \( N, n \) and \( c \) also. The quartiles show a stability in
the spacing of the values.

Analysis of Table 13 shows that the values are greater
than those from the normal population, which is to be

Table 13

Descriptive Statistics for Swaminathan's \( p \):
Negatively-skewed Population

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>( \sigma )</td>
</tr>
<tr>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>

Note. \( q_1 \) = skewness and \( q_2 \) = kurtosis
expected since in a negatively-skewed population more examinees would be classified as masters. The value of the index steadily increases with increasing sample size and number of items until it reaches $n=10, c=80\%$. At this point there is a slight decrease with increased $N$. As the criterion increases, however, the value of the index decreases. The standard deviation, and resulting range of values decreases regularly with increasing $N$ and increasing $n$, but increases as the criterion goes from 60\% to 80\%. The skewness values, as well as the kurtosis values, show a steady progress toward a normal curve when $c=80\%$, but they are more negative as $n$ increases when $c=60\%$. The mean is very near the median, noting that the increase of the value is greater with increase of $n$ for the higher criterion of 80\%.

Research Question 2e

What are the sampling distribution characteristics of Swaminathan's $k$, a threshold loss function, as used in determining the reliability of criterion-referenced tests?

The values of Swaminathan's $k$ are shown in Tables 14 and 15, as calculated from 1000 samples of a normally distributed and 1000 samples of a negatively-skewed population.

Looking at Table 14, it is obvious at first glance that these values are clearly lower than the values of Swaminathan's $p$ index. The same trends of a general increase in the index value with increase in sample size ($N$) and
number of item \(n\) continues. This increase occurs

Table 14

Descriptive Statistics for Swaminathan's \(k\):
Normal Population

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
<th>(q_1)</th>
<th>(q_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c)</td>
<td>(n)</td>
<td>(N)</td>
<td>(\mu)</td>
</tr>
<tr>
<td>60%</td>
<td>3</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>10</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>20</td>
<td>150</td>
</tr>
<tr>
<td>80%</td>
<td>4</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>10</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>20</td>
<td>150</td>
</tr>
</tbody>
</table>

Note. \(q_1\) = skewness and \(q_2\) = kurtosis
everywhere except when the number of items is \(n=10\) and \(n=20\)
with a criterion of 60%, at which time the values are
identical. There is a steady decrease in the standard deviation with increased \( N \), and a general decrease with increased \( n \), except when it moves from \( n=10 \) to \( n=20 \) with \( N=150 \) when the value stays the same. There is, however, generally a decrease in the value of Swaminathan's \( k \) index when the criterion is changed from a value of 60% to a value of 80%, as there is with the standard deviation and the range of values. The mean very closely approximates the median at nearly every point. The skewness and kurtosis show a rather erratic trend, with a general move toward more normal characteristics, as is evidenced by the smaller negative value for the skewness and a less positive value for the kurtosis. With a change in number of items or a change in the criterion, the trend is not as easily discerned.

Examination of Table 15, the negatively-skewed population values for Swaminathan's \( k \), reveals that these values are also very low for reliability values. The index value itself increases with increasing \( n \) and \( c \), and with \( N \) from \( n=10 \) to \( n=20 \), but decreases when item number goes from \( n=5 \) to \( n=10 \). The increase is fairly substantial from a criterion of 60% to a criterion of 80%. The standard deviation shows a fairly constant decrease under all conditions except from \( n=5 \) to \( n=10 \) with \( c=60\% \). The range, when \( N=30 \), is the greatest because negative values may be calculated for \( k \), causing the range to be greater than 1.0. The skewness and kurtosis values vary considerably, with a
frequent display of high positive values for the skewness.

The quartiles show a very unusual pattern; at \( N=30 \) one of
the quartiles is frequently 0. At \( c=60\% \), \( n=20 \), and \( N=150 \),
all quartiles show a value of 0.

Table 15

Descriptive Statistics for Swaminathan's \( k \):
Negatively-skewed Population

<table>
<thead>
<tr>
<th>Statistic</th>
<th>quartiles</th>
<th>( \hat{\psi}_1 )</th>
<th>( \hat{\psi}_2 )</th>
<th>(.25)</th>
<th>(.50)</th>
<th>(.75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>( n )</td>
<td>( N )</td>
<td>( \mu )</td>
<td>( \sigma )</td>
<td>range</td>
<td>( \hat{\psi}_1 )</td>
</tr>
<tr>
<td>60%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 5 30</td>
<td>.133</td>
<td>.279</td>
<td>1.13</td>
<td>1.606</td>
<td>1.460</td>
<td>-.03</td>
</tr>
<tr>
<td>3 5 150</td>
<td>.178</td>
<td>.148</td>
<td>.72</td>
<td>.041</td>
<td>-.491</td>
<td>.09</td>
</tr>
<tr>
<td>6 10 30</td>
<td>.102</td>
<td>.263</td>
<td>1.10</td>
<td>2.124</td>
<td>3.474</td>
<td>-.03</td>
</tr>
<tr>
<td>6 10 150</td>
<td>.155</td>
<td>.158</td>
<td>.83</td>
<td>.474</td>
<td>-.304</td>
<td>-.02</td>
</tr>
<tr>
<td>12 20 30</td>
<td>.134</td>
<td>.298</td>
<td>1.10</td>
<td>1.859</td>
<td>2.022</td>
<td>.00</td>
</tr>
<tr>
<td>12 20 150</td>
<td>.261</td>
<td>.210</td>
<td>1.04</td>
<td>.184</td>
<td>-.641</td>
<td>.00</td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 5 30</td>
<td>.292</td>
<td>.263</td>
<td>1.17</td>
<td>.047</td>
<td>-.403</td>
<td>.14</td>
</tr>
<tr>
<td>4 5 150</td>
<td>.297</td>
<td>.097</td>
<td>.70</td>
<td>-.039</td>
<td>.058</td>
<td>.23</td>
</tr>
<tr>
<td>8 10 30</td>
<td>.383</td>
<td>.218</td>
<td>1.22</td>
<td>-.207</td>
<td>-.207</td>
<td>.24</td>
</tr>
<tr>
<td>8 10 150</td>
<td>.395</td>
<td>.092</td>
<td>.54</td>
<td>-.014</td>
<td>-.108</td>
<td>.33</td>
</tr>
<tr>
<td>16 20 30</td>
<td>.261</td>
<td>.210</td>
<td>1.04</td>
<td>.184</td>
<td>-.641</td>
<td>.00</td>
</tr>
<tr>
<td>16 20 150</td>
<td>.461</td>
<td>.085</td>
<td>.54</td>
<td>.025</td>
<td>-.005</td>
<td>.40</td>
</tr>
</tbody>
</table>

Note. \( \hat{\psi}_1 \) = skewness and \( \hat{\psi}_2 \) = kurtosis
Research Question 3

How do the methods of Livingston, Brennan and Kane, Huynh, Subkoviak, and Swaminathan, compare with respect to their estimation of the proportion of consistent classifications (reliability) of the population?

After the population scores were simulated, a population value, $p$, the proportion of examinees correctly classified on parallel tests, was obtained. This parameter was calculated by adding the proportions of examinees that were correctly classified on both tests, for each of the varying conditions of $N$, $n$, and $c$ and was then used as a standard by which to compare the accuracy of the seven reliability indices studied. These results are summarized in Table 16 for the 1000 samples of $N=30$ each (a normal classroom size sample), from the normally distributed population.

Close examination of the threshold loss function, as used by Huynh, Subkoviak, and Swaminathan, listed in this table shows that, for this classroom-size sample of 30, Swaminathan's $p$ value gives a slightly positively biased estimate of the parameter value under all values of items ($n$) and criterion ($c$). Huynh's $p$ value gives slightly negatively biased estimates with the lower criterion of $c=60\%$, but slightly positively biased ones for the higher criterion of $c=80\%$. Subkoviak's values, for a criterion of 60%, exhibit estimates where as the number of items increases, the negative bias increases from approximately 15% to 20%
(calculated by dividing the deviation from the parameter by the parameter value). For a criterion of 80%, the negative bias increases from a low of around 60% for \( n=5 \) to a high of 85% for \( n=20 \).

Table 16

Means of Reliability Estimates: Normally distributed Population (\( N=30 \))

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Parameter P</th>
<th>( c = 60% )</th>
<th>( c = 80% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Livingston ( k^2/X )</td>
<td>.353</td>
<td>.531</td>
<td>.679</td>
</tr>
<tr>
<td>1</td>
<td>Brennan &amp; Kane M(C)</td>
<td>.353</td>
<td>.531</td>
<td>.679</td>
</tr>
<tr>
<td>2</td>
<td>Subkoviak ( p_c )</td>
<td>.546</td>
<td>.556</td>
<td>.530</td>
</tr>
<tr>
<td>2</td>
<td>Huynh p</td>
<td>.642</td>
<td>.685</td>
<td>.739</td>
</tr>
<tr>
<td>2</td>
<td>Swaminathan p</td>
<td>.647</td>
<td>.700</td>
<td>.739</td>
</tr>
<tr>
<td>2</td>
<td>Huynh k</td>
<td>.214</td>
<td>.344</td>
<td>.466</td>
</tr>
<tr>
<td>2</td>
<td>Swaminathan k</td>
<td>.247</td>
<td>.314</td>
<td>.314</td>
</tr>
</tbody>
</table>

Note. Type 1 = Squared-error loss function
Type 2 = Threshold loss function

Swaminathan’s and Huynh’s \( k \) values cannot be compared to the other values since they have a factor in the formula.
to eliminate chance classification of mastery or nonmastery which, because it is in the denominator of the formula, will always result in lower estimates. In addition to this fact, according to the literature, the kappa value has a range of values from -1.00 to +1.00 whereas the other indices have a range of values from 0.00 to +1.00. Of the two k values, then, Swaminathan's values are less negatively biased than Huynh's k at n=5, but increase less dramatically than Huynh's as n increases. At greater n, then, Huynh's k exhibits less negative bias than Swaminathan's k. With increasing criterion, both methods produce values which show a greater degree of negative bias.

When comparing the squared-error loss functions to each other, it becomes apparent that for this normally distributed population, $M(C)$ and $k^2(X, I_x)$ are measuring exactly the same thing. It is also noted that the estimates for a higher criterion of 80% are far superior to the estimates at c=60%.

In comparing the squared-error loss functions to the threshold loss functions as procedures for estimating the reliability, it seems that the squared-error estimates have much greater negative bias than the p value estimates of Huynh and Swaminathan for the lower criterion, and, that as n increases with the higher criterion, the values for $M(C)$ and $k^2(X, I_x)$ show a higher degree of positive bias than either of the two p values.
In Table 17, the mean values for the 1000 samples of larger sample size, \(N=150\), which is the average pupil load of a secondary teacher.

Examination of Table 17 reveals that, of the estimates that use the threshold loss function, the values of

Table 17

| Meas of Reliability Estimates: Normally distributed Population (\(N=150\)) |
|------------------|------------------|------------------|------------------|
| \(c = 60\%\) | \(c = 80\%\) |
| \(n =\) | 5 | 10 | 20 | 5 | 10 | 20 |
| Type | Method | Parameter \(\beta\) | .646 | .700 | .738 | .652 | .747 | .855 |
| 1 | Livingston \(k^{2}(X,T_{X})\) | .362 | .535 | .686 | .618 | .769 | .855 |
| 1 | Brennan & Kane \(M(C)\) | .362 | .535 | .686 | .618 | .769 | .855 |
| 2 | Subkoviak \(\hat{P}_{c}\) | .557 | .560 | .535 | .244 | .186 | .132 |
| 2 | Huynh \(P\) | .641 | .683 | .741 | .661 | .771 | .851 |
| 2 | Swaminathan \(P\) | .646 | .701 | .739 | .653 | .747 | .855 |
| 2 | Huynh \(k\) | .228 | .354 | .479 | .244 | .325 | .422 |
| 2 | Swaminathan \(k\) | .252 | .328 | .328 | .243 | .337 | .337 |

Note. Type 1 = Squared-error loss function  
Type 2 = Threshold loss function

Subkoviak's \(\hat{P}_{c}\) are still highly negatively biased, but not
as much as they are with a smaller sample size, except for $n=10$ and $n=20$ and $c=80\%$, in which case the estimates are even lower. The $p$ estimates of Huynh and Swaminathan continue to have less bias than Subkoviak, with Huynh's being slightly more positively biased than Swaminathan, which is expected since Swaminathan's estimate is calculated from two test results and Huynh's from one. The $k$ values of Swaminathan and Huynh follow the same general pattern: Swaminathan's $k$ is a better estimator at higher cutoff scores than is Huynh's $k$, but the reverse is true at the lower criterion.

As the two estimates that use the squared-error loss function are compared at this sample size, again the values for each are identical. These values exhibit an 80% negative bias for $n=5$ and $c=60\%$, decreasing to a 10% bias with increased $n$. For a criterion of 80%, the bias ranges from an 8% negative bias at $n=5$ to a 1% positive bias at $n=20$.

In comparing the squared-error loss function to the threshold loss function methods, all methods that use the threshold loss function (other than $k$), are less biased at the lower criterion; all except Subkoviak are also less biased at the higher criterion.

Tables 18 and 19 give the comparisons for the 1000 samples taken from the negatively-skewed population, with the former showing the results for the $N=30$ classroom sample size and the latter the results for the $N=150$ pupil load.
First examining the estimates employing the threshold loss function in Table 18 reveals that Subkoviak's estimates are highly biased: a 60% negative bias for \( n=5 \) to a 30% bias for \( n=20 \) when the criterion is 60%. When the criterion of 80% is considered, the negative bias decreases from 65% to 36% with an increase in \( n \). The \( p \) values of Huynh and
Swaminathan provide far better estimates of the parameter value with Huynh showing a decrease from 4% to 2% positive bias with increase in \( n \) for a criterion of 60%. For the higher criterion of 80%, the values are very slightly negatively biased until \( n=20 \), then there the bias is again positive. In considering the two values of \( k \), which are more negatively biased because of the chance factor being eliminated, Swaminathan's two test methods provide the more accurate estimates of the parameter, even though these values are still very negatively biased. For both values of \( c \), the estimates get larger (less negative bias) with increase on \( n \), but the estimates still tend to have as high as 89% negative bias with a low of around 36% with the highest \( n \).

Of the two estimates that use the squared-error loss function as part of their reliability indices, the values are again identical, giving further support to the notion that these two indices are measuring the same thing. The bias of these estimates falls between that of Subkoviak's \( P_c \) and Huynh's and Swaminathan's \( p \) values, exhibiting more negative bias than the latter and less negative bias than the former.

Further analysis of the negatively-skewed population is revealed by Table 19, which includes the results of the 1000 samples of \( N=150 \), taken, with replacement, from the population of 2000.

Examining the indices that use the threshold loss function, Subkoviak's \( P_c \) again gives the very negatively biased
estimates: ranging from 60% at n=5 to 30% at n=20. Whereas these estimates are fairly erratic with N=30, they are very consistent with N=150, showing a steady increase in the value

Table 19
Means of Reliability Estimates: Negatively-skewed Population (N=150)

<table>
<thead>
<tr>
<th>Type Method</th>
<th>Parameter P</th>
<th>n=5</th>
<th>10</th>
<th>20</th>
<th>c=60%</th>
<th>n=5</th>
<th>10</th>
<th>20</th>
<th>c=80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Livingston k^2(x, t, x)</td>
<td>0.858</td>
<td>.920</td>
<td>.959</td>
<td></td>
<td>0.549</td>
<td>.719</td>
<td>.836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brennan &amp; Kane M(C)</td>
<td>0.858</td>
<td>.920</td>
<td>.959</td>
<td></td>
<td>0.549</td>
<td>.719</td>
<td>.836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subkoviak P_c</td>
<td>0.382</td>
<td>.545</td>
<td>.679</td>
<td></td>
<td>0.294</td>
<td>.432</td>
<td>.538</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huynh p</td>
<td>0.975</td>
<td>.981</td>
<td>.989</td>
<td></td>
<td>0.816</td>
<td>.821</td>
<td>.852</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swaminathan p</td>
<td>0.948</td>
<td>.961</td>
<td>.978</td>
<td></td>
<td>0.828</td>
<td>.831</td>
<td>.841</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huynh k</td>
<td>0.107</td>
<td>.200</td>
<td>.294</td>
<td></td>
<td>0.225</td>
<td>.375</td>
<td>.501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swaminathan k</td>
<td>0.178</td>
<td>.155</td>
<td>.261</td>
<td></td>
<td>0.297</td>
<td>.395</td>
<td>.461</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Type 1 = Squared-error loss function
Type 2 = Threshold loss function

of the estimate (decrease in negative bias). Huynh's and Swaminathan's p estimates are again closer to the parameter value, with Huynh's p estimate exhibiting a much greater
degree of positive bias than do Swaminathan's estimates. Both estimates show a steady increase in the value of $p$ with increasing $n$, but a decrease from $c=60\%$ to $c=80\%$.

Upon consideration of the two kappa ($k$) estimates, it is noted that these are once again considerably lower, due to the chance element placed in the denominator of the formula. The estimates made by Swaminathan's $k$ are, as expected, less biased than are Huynh's. Both values of $k$ are much closer to the parameter value with a higher criterion, as calculated from this negatively-skewed population.

The two indices which use the squared-error loss function again have identical values. These values are negatively biased throughout the table, but are less so when $n$ is increased. These values fall closer to Huynh's $p$ and Swaminathan's $p$ than previously, and are superior to Subkoviak's values.

**Research Question 4**

How are the indices of reliability which use the threshold loss and squared-error loss function compatible with the condition of preferring to use two cutoff scores with a zone of indifference in the middle?

There are instances when one would only be concerned with identifying those who are truly masters and those who are truly nonmasters, with those falling in the middle of these two regions being of little concern to the examiner at
that time. For an examinee with a true score in the indifference zone, it does not matter whether he is classified as a master or nonmaster (Huynh, 1980a). This situation would obviously arise only in situations where the decisions are not concerning life-threatening situations.

To determine which of these indices would be most compatible under these conditions, the estimates were examined to discern which one gives the most clearcut accuracy at both the 60% cutoff score level and the 80% level. The ideal situation would be one that gives the most unbiased estimates at both cutoff points, regardless of \( n \).

Table 20 shows the estimates from each procedure with \( N=150 \) and \( n=10 \), and from the normally distributed

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Parameter</th>
<th>( \theta = 60% )</th>
<th>Parameter</th>
<th>( \theta = 80% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brennan &amp; Kane*</td>
<td>.700</td>
<td>.535</td>
<td>.747</td>
<td>.769</td>
</tr>
<tr>
<td>Livingston*</td>
<td>.700</td>
<td>.535</td>
<td>.747</td>
<td>.769</td>
</tr>
<tr>
<td>Subkoviak</td>
<td>.700</td>
<td>.560</td>
<td>.747</td>
<td>.186</td>
</tr>
<tr>
<td>Huynh ( p )</td>
<td>.700</td>
<td>.683</td>
<td>.747</td>
<td>.771</td>
</tr>
<tr>
<td>Swaminathan ( p )</td>
<td>.700</td>
<td>.701</td>
<td>.747</td>
<td>.747</td>
</tr>
<tr>
<td>Huynh ( k )</td>
<td>.700</td>
<td>.354</td>
<td>.747</td>
<td>.325</td>
</tr>
<tr>
<td>Swaminathan ( k )</td>
<td>.700</td>
<td>.314</td>
<td>.747</td>
<td>.322</td>
</tr>
</tbody>
</table>

* use the squared-error loss function
population. The index that has the greatest accuracy of estimation at both points is Swaminathan's $p$ index, showing a very slight positive bias at both points. Huynh's $p$ value is also very close to the parameter value at both points, and if two testings are not feasible, would be a good choice. If the element of chance is to be considered, so as to be more certain that those classified below 60% are true nonmasters and those above 80% are true masters, then Swaminathan's $k$ value appears to be the estimate of choice.

Research Question 5

How are the different reliability indices, with regard to either squared-error loss or threshold loss, compatible under conditions where losses related to all false mastery and false nonmastery decision errors are not equally serious?

In Tables 16 to 19, the calculated values for the different methods that use the squared-error loss function and the threshold loss function are displayed. The statistics speak for themselves as to which shows less bias: the squared-error loss methods, while reasonable, are not the most reliable with regard to assigning correct mastery states based on the results of a test. They are consistently lower, $p$ values, as well as Subkoviak's $P_c$ when $c=60\%$ for the normal distribution; otherwise, they are less negatively biased than the $P_c$. They do display less negative bias than Huynh's $k$ or Swaminathan's $k$, but then, those values
have a different range (-1 to +1) and cannot be directly compared.


CHAPTER 5

FINDINGS, CONCLUSIONS, RECOMMENDATIONS

This study was undertaken in order to compare the reliability indices of criterion-referenced tests and their compatibility with specific conditions of cutoff scores, while at the same time examining the sampling characteristics of each index. In order to eliminate any extraneous factors that might affect test scores, data were randomly generated by SPSSX which resulted in three parallel test forms with five, ten, and twenty items each for two populations of 2000 each: normally distributed and negatively-skewed. A computer program was then created by GAUSS that would make 1000 random samples, with replacement, from each population and then calculate each of seven reliability indices for these samples. A mean value for these 1000 calculations was then reported, along with the descriptive statistics for each set of samples.

The data were then organized so that one could analyze the patterns of the descriptive statistics for each of the seven methods of computing the reliability and thereby determine the sampling distribution characteristics of each. In addition, the data were also organized so that one could compare the seven methods and determine the efficiency of each.
Findings

Research Questions 1 & 2

What are the sampling characteristics of the seven methods of determining the reliability of criterion-referenced tests?

According to the central limits theorem, the sampling distribution of means will approach a normal distribution of means as sample size increases, despite departures from normality in the population distributions. In every instance of the 1000 sample sets drawn from the populations of 2000, the mean for that index estimated the population value for that index \( \pm 0.02 \). In addition the skewness values for all seven indices that were calculated, from both populations, more closely approximated characteristics of a normal curve as sample size increased. This was true if it was an index that made use of either the squared-error loss function or the threshold loss function. Some of these values, such as those found in Brennan and Kane's \( M(C) \) and Livingston's \( k_x^2(X, I_x) \) were very negatively skewed and very leptokurtic while others, such as Subkoviak's \( P_C \), had values that closely approximated normal curve characteristics. Both Swaminathan's and Huynh's values for both \( p \) and \( k \) were slightly negatively skewed and slightly leptokurtic. Very few samples exhibited a positively skewed value or a platykurtic value.
Research Question 3

How do the methods of Livingston, Brennan and Kane, Huynh, Subkoviak, and Swaminathan compare with respect to their estimation of the proportion of consistent classifications (reliability) of the population?

The $k^2(X, I_x)$ and $M(C)$ squared-error function indices proposed by Livingston and Brennan and Kane respectively proved to measure reliability identically, even though the actual formulas appear to be very different. Both of these indices have no correction for chance, have no distributional assumptions, have values that increase as the cutoff score is set farther from the mean, and index values that increase as the number of items increases. The fact that the value increases as the difference between the mean and the cutoff score increases appears to make this test a function of the cutoff score as well as a function of the individual's response to the items, which in turn means that it is not really a reliability coefficient since it is not directly related to the repetition of the measure. Instead, it seems to measure the consistency of scores in relation to the cutoff score in repeated testings. Livingston's thoughts concerning this issue are clear: "...this index deserves to be called a reliability coefficient because it represents the ratio of 'true' to 'observed' mean squared deviations from the criterion score, and the criterion-referenced correlation between alternative forms of the same test, and the squared
criterion-referenced correlation between true scores and observed scores" (Livingston, 1972, p. 139).

Of the indices studied which make use of the threshold loss function, Subkoviak's index gives the poorest estimates of the parameter value. The reasons for this appear to be based in the assumptions Subkoviak makes for the use of his index: the simple binomial model and independent scores. A very basic part of his formula is the calculation of $P_c$, based on the assumption that person $i$ is a member of a group with a unimodal distribution of scores. It has been stated in the literature that "grossly inaccurate estimates may occur if the linear regression approximations of binomial parameter $P_c$ are blindly observed for multimodal data sets" (Subkoviak, 1978a, p. 268). Even though the population follows the simple binomial model, the separate sample populations frequently show very different means and variance than the population itself. When the sample populations were examined closely, and frequency distribution graphs drawn, it became very clear that, although the populations from which the samples were drawn were unimodal, the samples were frequently bimodal and occasionally even trimodal. For this reason, Subkoviak's $P_c$ was found to result in very negatively biased estimates.

The last four methods of calculating criterion-referenced reliability, as proposed by Huynh and Swaminathan, result in $p$ and $k$ values, both of which make use of the
threshold loss function. Swaminathan's estimates, as well as Huynh's estimates, are very accurate but Swaminathan's are based on two test administrations whereas Huynh's is calculated from one administration. Generally, the standard error associated with a single administration procedure is smaller than with two administrations since Huynh's method assumes that the data follow the beta-binomial model and Swaminathan's method does not. Therefore, when the test data reflects the beta-binomial model, the single-administration estimates will be based on more complete information and would therefore be more accurate.

From a statistical point of view, the single test administration is a parametric estimation and the two test administration estimates are nonparametric. Parametric procedures are more powerful and more accurate if the underlying model is correct (Huynh and Saunders, 1980).

The comparison between \( k \) and \( p \) shows that there is a larger bias for \( k \) as compared to \( p \), which is to be expected with the factor \( \frac{1-P}{c} \) (which cannot exceed .50) in the denominator of the equation which defines \( k \). The bias of \( k \) is at least twice that associated with \( p \) every time, usually much more. The approximations used to calculate Huynh's \( p \) and \( k \) have been shown to provide a 10% bias for \( k \) and an 1% bias for \( p \). For the negatively-skewed populations, situations where a high proportion of examinees will be classified as masters, \( 1-P_c \) are close to zero, resulting in a very high
bias for the $k$ value. The value of $k$ seems to increase with
increase $c$, but after a maximum value, the value decreases.
This observation can be partly explained by noting that $p$ is
near 1 when the cutoff score is too small or too large. For
these extreme cases, there does not appear to be much room
for improvement of consistency of decisions beyond that of
chance.

**Research Question 4**

how are the indices of reliability which use the
threshold loss and squared-error loss function compatible
with the condition of preferring to use two cutoff scores
with a zone of indifference in the middle?

When there is a zone of indifference between mastery
classification and nonmastery classification where it does
not really matter which way an examinee is classified, then
examination of the data reveals that either Huynh's $p$ or
Swaminathan's $p$ would closely approximate the correct
classifications at both a low cutoff score and a high cutoff
score. Both of the squared-error loss indices are reasonably
accurate for this also, with the lower cutoff region showing
the most negative bias. If the need is to estimate the
proportion of correct classifications without regard to
chance, then Huynh's $k$ value is the better of the two for
both a low cutoff point and a high one with a zone of
indifference in the middle.
Research Question 5

How are the different reliability indices, with regard to either squared-error loss or threshold loss, compatible under conditions where losses related to all false mastery and false nonmastery decision errors are not equally serious?

Findings were rather inconclusive with regard to this question in that to actually decide which of these values is better under the conditions described above, one would need to study actual data in greater detail. From the scope of this investigation, the "eyeball" results show that, if one assumes that false mastery decisions are more serious one would want the most accurate estimator, as one would if false nonmastery decisions were more serious. With only those criteria as guidelines, the findings would indicate that, again the Huynh or Swaminathan $p$ would be the most useful if those classifications made through chance are of no concern; if they are, then the Huynh $k$ is the more accurate index that takes chance into consideration when calculated, although it is a very poor reliability index.

Conclusions

The sampling distribution characteristics of the two indices that use the squared-error loss function and the five indices that use the threshold loss function show that as the sample size increases, the mean value of each index more closely approximate the normal curve characteristics in that they become less negatively-skewed and less leptokurtic.
Of the seven methods evaluated, Subkoviak's $P_C$ has characteristics most like the normal curve.

Livingston's $k^2(X,T)$ and Brennan and Kane's $M(C)$ methods of computing the reliability of a criterion-referenced test cannot be used or interpreted like conventional reliability indices because they do not estimate the ratio of true score variance to observed variance alone, but instead they are sensitive to the difference between the cutoff score and the mean rather than the consistently with which an examinee is assigned to a mastery state. Reliability in criterion-referenced testing is the consistency of the decisions concerning mastery status, and not how far an examinee's score deviates from the cutoff score. For this reason, the conclusion is that the squared-error loss indices should not be used in determining classifications consistency.

Subkoviak's index has very strict assumptions and if these assumptions are not followed, the index will result in grossly inaccurate estimates. Specifically, if one cannot ascertain for certain that the sampling distribution is unimodal, Subkoviak's index should not be used.

The overall conclusion concerning which index provides the best estimate of the parameter value of classification consistency (reliability) is that Huynh's $p$ is the statistic of choice. It is considered better than Swaminathan's $p$ because Huynh's $p$ is a parametric statistic and therefore will provide the more accurate estimate of the parameter.
value of classification consistency. This same $p$ statistic, as calculated by Huynh's method, also is the choice if one wishes to use two criteria, with a zone of indifference in the middle, as it is if the losses related to all false mastery and false nonmastery decision errors are not equally serious. The $k$ indices provide such poor estimates of reliability that they are not recommended for use alone. In combination with a $p$ value, they could provide additional information concerning the classification consistency beyond that expected by chance.

Recommendations

The purpose of the present study was to examine the various reliability indices used in criterion-referenced testing. Hopefully the information gained from this investigative study will assist test-users and test-makers of criterion-referenced instruments in making more informed decisions about which index to use under what conditions. With that in mind, several recommendations have been included for use of these indices or for further investigation which came out of this study or were beyond the scope of this study. These recommendations are as follows:

1. The choice of index should be based on all of the following: test form assumptions, position of the cutoff score, intended score interpretation, the type of decision to be made from the test, and the underlying model assumptions.
2. All criterion-referenced tests should include reliability data and the method by which it was obtained.

3. An easy-to-use computer program that would estimate reliability using Huynh's very complex formula would be very beneficial to criterion-referenced test users.

4. There is a need for an additional index that can be used when losses that are not equally serious are of concern.
CHAPTER BIBLIOGRAPHY


APPENDIX A

SPSS PROGRAM TO GENERATE RANDOM DATA

SAMPLE OF RANDOM DATA
CREATING DATA BY RANDOM NUMBER

INPUT PROGRAM
LOOP B1 = 1 TO 2000
COMPUTE X1 = RND(UNIFORM(99)/10 + .5)
COMPUTE X2 = RND(UNIFORM(99)/10 + .5)
COMPUTE X3 = RND(UNIFORM(99)/10 + .5)
COMPUTE X4 = RND(UNIFORM(99)/10 + .5)
COMPUTE X5 = RND(UNIFORM(99)/10 + .5)
COMPUTE X6 = RND(UNIFORM(99)/10 + .5)
COMPUTE X7 = RND(UNIFORM(99)/10 + .5)
COMPUTE X8 = RND(UNIFORM(99)/10 + .5)
COMPUTE X9 = RND(UNIFORM(99)/10 + .5)
COMPUTE X10 = RND(UNIFORM(99)/10 + .5)
COMPUTE X11 = RND(UNIFORM(99)/10 + .5)
COMPUTE X12 = RND(UNIFORM(99)/10 + .5)
COMPUTE X13 = RND(UNIFORM(99)/10 + .5)
COMPUTE X14 = RND(UNIFORM(99)/10 + .5)
COMPUTE X15 = RND(UNIFORM(99)/10 + .5)
COMPUTE X16 = RND(UNIFORM(99)/10 + .5)
COMPUTE X17 = RND(UNIFORM(99)/10 + .5)
COMPUTE X18 = RND(UNIFORM(99)/10 + .5)
COMPUTE X19 = RND(UNIFORM(99)/10 + .5)
COMPUTE X20 = RND(UNIFORM(99)/10 + .5)
COMPUTE X21 = RND(UNIFORM(99)/10 + .5)
COMPUTE X22 = RND(UNIFORM(99)/10 + .5)
COMPUTE X23 = RND(UNIFORM(99)/10 + .5)
COMPUTE X24 = RND(UNIFORM(99)/10 + .5)
COMPUTE X25 = RND(UNIFORM(99)/10 + .5)
COMPUTE X26 = RND(UNIFORM(99)/10 + .5)
COMPUTE X27 = RND(UNIFORM(99)/10 + .5)
COMPUTE X28 = RND(UNIFORM(99)/10 + .5)
COMPUTE X29 = RND(UNIFORM(99)/10 + .5)
COMPUTE X30 = RND(UNIFORM(99)/10 + .5)
COMPUTE X31 = RND(UNIFORM(99)/10 + .5)
COMPUTE X32 = RND(UNIFORM(99)/10 + .5)
COMPUTE X33 = RND(UNIFORM(99)/10 + .5)
COMPUTE X34 = RND(UNIFORM(99)/10 + .5)
COMPUTE X35 = RND(UNIFORM(99)/10 + .5)
COMPUTE X36 = RND(UNIFORM(99)/10 + .5)
COMPUTE X37 = RND(UNIFORM(99)/10 + .5)
COMPUTE X38 = RND(UNIFORM(99)/10 + .5)
COMPUTE X39 = RND(UNIFORM(99)/10 + .5)
COMPUTE X40 = RND(UNIFORM(99)/10 + .5)
COMPUTE X41 = RND(UNIFORM(99)/10 + .5)
COMPUTE X42 = RND(UNIFORM(99)/10 + .5)
COMPUTE X43 = RND(UNIFORM(99)/10 + .5)
COMPUTE X44 = RND(UNIFORM(99)/10 + .5)
COMPUTE X45 = RND(UNIFORM(99)/10 + .5)
COMPUTE X46 = RND(UNIFORM(99)/10 + .5)
COMPUTE X47 = RND(UNIFORM(99)/10 + .5)
COMPUTE X48 = RND(UNIFORM(99)/10 + .5)
COMPUTE X49 = RND(UNIFORM(99)/10 + .5)
COMPUTE X50 = RND(UNIFORM(99)/10 + .5)
COMPUTE X51 = RND(UNIFORM(99)/10 + .5)
COMPUTE X52 = RND(UNIFORM(99)/10 + .5)
COMPUTE X53 = RND(UNIFORM(99)/10 + .5)
COMPUTE X54 = RND(UNIFORM(99)/10 + .5)
COMPUTE X55 = RND(UNIFORM(99)/10 + .5)
COMPUTE X56 = RND(UNIFORM(99)/10 + .5)
COMPUTE X57 = RND(UNIFORM(99)/10 + .5)
COMPUTE X58 = RND(UNIFORM(99)/10 + .5)
COMPUTE X59 = RND(UNIFORM(99)/10 + .5)
COMPUTE X60 = RND(UNIFORM(99)/10 + .5)
END LOOP
END CASE
END LOOP
END FILE
FORMAT XI TO X60 (F2.0)
LIST VARS = XI TO X60
### Sample of Random Data Generated by SPSSX

| 7 | 3 | 2 | 7 | 4 | 5 | 1 | 2 | 2 | 1 | 7 | 2 | 8 | 10 | 6 | 3 | 8 | 1 | 1 | 0 | 3 | 7 | 10 | 3 | 4 | 1 | 8 | 2 | 1 | 2 | 1 | 7 | 7 | 9 | 4 | 6 | 7 |
| 6 | 7 | 10 | 4 | 8 | 2 | 6 | 10 | 4 | 6 | 3 | 1 | 10 | 5 | 1 | 4 | 3 | 8 | 9 | 2 | 2 | 7 | 2 | 4 | 6 | 6 | 7 | 1 | 5 | 2 | 1 | 0 | 6 | 3 | 7 | 7 | 9 | 1 | 8 | 9 |
| 7 | 9 | 3 | 2 | 4 | 5 | 1 | 8 | 1 | 2 | 3 | 7 | 8 | 4 | 7 | 10 | 4 | 3 | 2 | 2 | 7 | 2 | 4 | 6 | 8 | 1 | 6 | 5 | 8 | 1 | 6 | 1 | 0 | 2 | 4 | 5 | 3 | 2 | 7 |
| 10 | 3 | 7 | 7 | 6 | 10 | 4 | 10 | 6 | 4 | 8 | 4 | 5 | 1 | 2 | 8 | 10 | 3 | 7 | 1 | 1 | 0 | 9 | 6 | 2 | 7 | 5 | 5 | 5 | 3 | 1 | 6 | 1 | 4 | 8 | 6 | 1 | 4 |
| 8 | 4 | 6 | 7 | 4 | 5 | 6 | 1 | 5 | 1 | 1 | 9 | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9 | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 10 | 7 | 6 | 1 | 2 | 10 | 9 | 9 | 3 | 7 | 4 | 10 | 3 | 9 | 3 | 1 | 2 | 4 | 3 | 3 | 9 | 6 | 2 | 2 | 10 | 1 | 5 | 10 | 5 | 9 | 8 | 5 | 2 | 3 | 2 | 7 |
| 10 | 9 | 9 | 7 | 6 | 7 | 2 | 1 | 8 | 6 | 10 | 8 | 5 | 5 | 8 | 1 | 4 | 8 | 5 | 4 | 8 | 3 | 5 | 1 | 2 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 |
| 8 | 6 | 3 | 1 | 1 | 9 | 3 | 8 | 4 | 9 | 7 | 2 | 4 | 8 | 5 | 4 | 8 | 2 | 3 | 0 | 9 | 7 | 9 | 4 | 6 | 1 | 2 | 1 | 0 | 9 | 6 | 2 | 7 | 5 | 5 | 8 | 1 | 4 | 8 |
| 2 | 8 | 0 | 4 | 6 | 2 | 8 | 3 | 1 | 3 | 9 | 10 | 0 | 1 | 2 | 2 | 2 | 10 | 9 | 3 | 4 | 9 | 10 | 6 | 3 | 9 | 8 | 5 | 1 | 4 | 1 | 5 | 2 | 1 | 1 | 7 | 10 |
| 9 | 4 | 7 | 7 | 6 | 2 | 4 | 1 | 0 | 3 | 5 | 10 | 1 | 3 | 7 | 6 | 7 | 4 | 5 | 5 | 1 | 9 | 7 | 9 | 4 | 6 | 1 | 4 | 8 | 5 | 4 | 8 | 3 | 5 | 1 | 2 | 2 | 3 |
| 8 | 2 | 9 | 3 | 5 | 7 | 10 | 1 | 4 | 9 | 7 | 9 | 9 | 10 | 8 | 4 | 2 | 7 | 9 | 5 | 10 | 9 | 0 | 2 | 1 | 3 | 10 | 1 | 5 | 5 | 1 | 3 | 8 | 1 | 10 | 2 | 8 |
| 8 | 6 | 7 | 6 | 2 | 10 | 1 | 8 | 6 | 3 | 1 | 5 | 1 | 6 | 9 | 6 | 7 | 4 | 6 | 5 | 1 | 2 | 2 | 3 | 1 | 2 | 0 | 5 | 8 | 9 | 6 | 10 | 9 | 4 | 6 | 8 | 7 | 5 |
| 8 | 6 | 5 | 7 | 10 | 6 | 1 | 7 | 2 | 10 | 1 | 1 | 5 | 1 | 6 | 5 | 8 | 9 | 6 | 10 | 9 | 4 | 6 | 8 | 7 | 5 | 1 | 0 | 5 | 10 | 5 | 2 | 3 | 5 | 7 | 5 | 6 |
| 9 | 5 | 3 | 4 | 8 | 9 | 10 | 4 | 5 | 7 | 7 | 8 | 7 | 9 | 5 | 10 | 1 | 3 | 7 | 4 | 10 | 6 | 9 | 10 | 6 | 1 | 8 | 3 | 1 | 4 | 10 | 9 | 10 | 6 | 8 | 7 | 5 |
| 9 | 3 | 7 | 6 | 2 | 9 | 6 | 5 | 5 | 3 | 6 | 10 | 7 | 10 | 14 | 6 | 7 | 4 | 6 | 5 | 4 | 8 | 5 | 10 | 9 | 4 | 6 | 8 | 7 | 5 | 1 | 0 | 5 | 10 | 5 | 2 | 3 |
| 9 | 4 | 7 | 7 | 4 | 3 | 5 | 5 | 7 | 5 | 10 | 9 | 3 | 8 | 6 | 10 | 9 | 8 | 4 | 2 | 4 | 5 | 4 | 7 | 5 | 6 | 7 | 5 | 9 | 4 | 10 | 3 | 2 | 4 | 5 | 6 | 7 |
| 9 | 8 | 3 | 7 | 4 | 5 | 7 | 8 | 5 | 10 | 4 | 10 | 9 | 2 | 8 | 4 | 3 | 7 | 4 | 6 | 4 | 4 | 3 | 1 | 9 | 6 | 7 | 8 | 4 | 5 | 6 | 7 | 8 | 4 | 3 | 7 | 4 | 6 |
| 5 | 7 | 2 | 6 | 10 | 0 | 3 | 10 | 8 | 5 | 8 | 8 | 7 | 1 | 4 | 1 | 6 | 1 | 1 | 5 | 3 | 5 | 4 | 9 | 2 | 3 | 6 | 6 | 7 | 2 | 2 | 5 | 3 | 10 | 1 | 5 | 2 |
| 3 | 8 | 5 | 6 | 5 | 3 | 1 | 6 | 8 | 2 | 7 | 6 | 1 | 7 | 3 | 6 | 9 | 6 | 10 | 3 | 7 | 6 | 9 | 8 | 5 | 6 | 7 | 8 | 9 | 2 | 4 | 3 | 7 | 5 | 7 | 7 | 6 |
| 5 | 10 | 8 | 4 | 6 | 10 | 9 | 5 | 7 | 4 | 8 | 4 | 8 | 4 | 7 | 4 | 0 | 6 | 4 | 1 | 7 | 1 | 2 | 7 | 7 | 5 | 3 | 10 | 6 | 6 | 9 | 2 | 4 | 3 | 7 | 5 | 7 | 7 |
| 8 | 1 | 7 | 2 | 4 | 8 | 10 | 9 | 9 | 4 | 9 | 2 | 4 | 2 | 10 | 4 | 9 | 2 | 4 | 2 | 10 | 4 | 9 | 2 | 4 | 2 | 10 | 4 | 9 | 2 | 4 | 2 | 10 | 4 | 9 | 2 | 4 |
| 2 | 0 | 4 | 3 | 8 | 7 | 6 | 2 | 8 | 1 | 1 | 1 | 7 | 4 | 4 | 4 | 8 | 9 | 4 | 10 | 2 | 5 | 7 | 1 | 6 | 1 | 3 | 10 | 5 | 10 | 4 | 7 | 9 | 1 | 2 |
| 4 | 0 | 7 | 7 | 8 | 7 | 2 | 1 | 7 | 3 | 2 | 10 | 7 | 9 | 2 | 2 | 2 | 10 | 9 | 8 | 9 | 9 | 9 | 8 | 9 | 3 | 4 | 2 | 4 | 2 | 8 | 3 | 0 | 10 | 2 | 2 | 1 | 2 | 1 | 5 | 3 |
| 5 | 6 | 1 | 9 | 3 | 5 | 8 | 7 | 6 | 4 | 5 | 7 | 3 | 9 | 4 | 10 | 1 | 4 | 8 | 5 | 4 | 8 | 3 | 5 | 1 | 2 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 |

93
APPENDIX B

GENERATION OF POPULATION PROBABILITIES
1. Computer Program to Generate Population Probabilities: SAS

DATA A;
RETAIN SEED 1 35745924;
DO I = 1 TO 2000;
XI = RANBIN(SEED1, 10, .5);
OUTPUT;
END;
PROC FREQ;
TABLES XI;
DATA B;
RETAIN SEED2 35745924;
DO I = 1 TO 2000;
X2 = RANBIN(SEED2, 10, .8);
OUTPUT;
END;
PROC FREQ;
TABLES X2;

2. Output from above SAS program

<table>
<thead>
<tr>
<th>X1</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>254</td>
</tr>
<tr>
<td>4</td>
<td>428</td>
</tr>
<tr>
<td>5</td>
<td>476</td>
</tr>
<tr>
<td>6</td>
<td>426</td>
</tr>
<tr>
<td>7</td>
<td>210</td>
</tr>
<tr>
<td>8</td>
<td>74</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X2</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>172</td>
</tr>
<tr>
<td>7</td>
<td>408</td>
</tr>
<tr>
<td>8</td>
<td>582</td>
</tr>
<tr>
<td>9</td>
<td>564</td>
</tr>
<tr>
<td>10</td>
<td>220</td>
</tr>
</tbody>
</table>
APPENDIX C

RECODING PROGRAM AND SAMPLE DATA
1. Sample recoding program for N=1000 from SPSSX

DO IF (CASENUM LE 17)
RECODE IT1 TO IT60 (10=1) (11=1) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 17 AND CASENUM LT 69)
RECODE IT1 TO IT60 (10=1) (11=1) (12=1) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 70 AND CASENUM LT 180)
RECODE IT1 TO IT60 (10=1) (11=2) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 181 AND CASENUM LT 394)
RECODE IT1 TO IT60 (10=1) (11=3) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 395 AND CASENUM LT 632)
RECODE IT1 TO IT60 (10=1) (11=4) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 633 AND CASENUM LT 645)
RECODE IT1 TO IT60 (10=1) (11=5) (12=1) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 646 AND CASENUM LT 950)
RECODE IT1 TO IT60 (10=1) (11=5) (12=2) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 951 AND CASENUM LT 975)
RECODE IT1 TO IT60 (10=1) (11=5) (12=3) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

DO IF (CASENUM GE 976 AND CASENUM LT 1000)
RECODE IT1 TO IT60 (10=1) (11=5) (12=4) (ELSE=0)
 COMPUTE SCIN5=SUM(IT1 TO IT5)
 COMPUTE SCIN5=SUM(IT6 TO IT10)
 COMPUTE SCIN10=SUM(IT11 TO IT15)
 COMPUTE SCIN10=SUM(IT16 TO IT20)
 COMPUTE SCIN20=SUM(IT21 TO IT25)
 COMPUTE SCIN20=SUM(IT26 TO IT30)
 COMPUTE SCIN30=SUM(IT31 TO IT35)
 COMPUTE SCIN30=SUM(IT36 TO IT40)
 END IF

END IF
1. Sample of recoded data (recoded from raw data in Appendix A)
APPENDIX D

SUMMATION OF SCORES
1. Sample of scores collected from COMPUTE SUM in program from Appendix C

<table>
<thead>
<tr>
<th>SCN1</th>
<th>SCN2</th>
<th>SCN10</th>
<th>SCN110</th>
<th>SCN120</th>
<th>SCN130</th>
<th>SCN140</th>
<th>SCN150</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>2.00</td>
<td>4.00</td>
<td>5.00</td>
<td>7.00</td>
<td>7.00</td>
</tr>
<tr>
<td>2.00</td>
<td>2.00</td>
<td>4.00</td>
<td>2.00</td>
<td>6.00</td>
<td>3.00</td>
<td>5.00</td>
<td>4.00</td>
</tr>
<tr>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>5.00</td>
<td>4.00</td>
</tr>
<tr>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td>6.00</td>
<td>5.00</td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>6.00</td>
<td>5.00</td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
</tr>
<tr>
<td>8.00</td>
<td>8.00</td>
<td>8.00</td>
<td>8.00</td>
<td>8.00</td>
<td>8.00</td>
<td>8.00</td>
<td>8.00</td>
</tr>
<tr>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
</tr>
<tr>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
</tr>
</tbody>
</table>
APPENDIX E

MATRICES USED TO COMPUTE PARAMETER VALUE, $P$
### 1. Matrices for Normally Distributed Population: n=5 and n=10

#### n=5

<table>
<thead>
<tr>
<th>COUNT</th>
<th>SC1NS</th>
<th>01</th>
<th>11</th>
<th>21</th>
<th>31</th>
<th>41</th>
<th>51</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>2</td>
<td>13</td>
<td>11</td>
<td>5</td>
<td>3</td>
<td></td>
<td>3.6</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>28</td>
<td>34</td>
<td>30</td>
<td>13</td>
<td>3</td>
<td>11.7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>5</td>
<td>32</td>
<td>60</td>
<td>53</td>
<td>40</td>
<td>18</td>
<td>21.7</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>10</td>
<td>29</td>
<td>66</td>
<td>70</td>
<td>75</td>
<td>23</td>
<td>27.6</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td>13</td>
<td>45</td>
<td>72</td>
<td>68</td>
<td>38</td>
<td>23.7</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>19</td>
<td>34</td>
<td>43</td>
<td>119</td>
</tr>
</tbody>
</table>

#### n=10

<table>
<thead>
<tr>
<th>COLUMN TOTAL</th>
<th>3.1</th>
<th>12.0</th>
<th>24.2</th>
<th>24.6</th>
<th>23.3</th>
<th>22.5</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW TOTAL</td>
<td>3.6</td>
<td>11.7</td>
<td>21.7</td>
<td>27.6</td>
<td>23.7</td>
<td>119</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**NUMBER OF MISSING OBSERVATIONS** = 0
3. Matrix for Normal Distribution: n=20

<table>
<thead>
<tr>
<th>COUNT</th>
<th>21</th>
<th>31</th>
<th>41</th>
<th>51</th>
<th>61</th>
<th>71</th>
<th>81</th>
<th>91</th>
<th>101</th>
<th>111</th>
<th>121</th>
<th>131</th>
<th>141</th>
<th>151</th>
<th>161</th>
<th>171</th>
<th>181</th>
<th>191</th>
<th>201</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

| COLUMN TOTAL | 10 | 7 | 27 | 33 | 40 | 47 | 54 | 61 | 68 | 75 | 82 | 89 | 96 | 103 | 110 | 117 | 124 | 131 | 138 |
|--------------|----|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| TOTAL        | 10 | 7 | 27 | 33 | 40 | 47 | 54 | 61 | 68 | 75 | 82 | 89 | 96 | 103 | 110 | 117 | 124 | 131 | 138 |

NUMBER OF MISSING OBSERVATIONS = 0
### 3. Matrices for Negatively-skewed Population: \( n=5 \) and \( n=10 \)

**SC1N5**

<table>
<thead>
<tr>
<th>COUNT</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC2N5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>.1</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>7</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>11</td>
<td>23</td>
<td>35</td>
<td>31</td>
<td>101</td>
<td>1001</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>13</td>
<td>44</td>
<td>75</td>
<td>89</td>
<td>222</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8</td>
<td>18</td>
<td>66</td>
<td>520</td>
<td>634</td>
<td>63.4</td>
</tr>
</tbody>
</table>

**SC1N10**

<table>
<thead>
<tr>
<th>COUNT</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC2N10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>23</td>
<td>20</td>
<td>19</td>
<td>16</td>
<td>9</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>23</td>
<td>20</td>
<td>19</td>
<td>16</td>
<td>9</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>22</td>
<td>19</td>
<td>5</td>
</tr>
</tbody>
</table>

**COLUMN TOTAL**

<table>
<thead>
<tr>
<th></th>
<th>.1</th>
<th>.4</th>
<th>9.9</th>
<th>29.8</th>
<th>64.8</th>
<th>100.0</th>
</tr>
</thead>
</table>

**NUMBER OF MISSING OBSERVATIONS** = 0
### 4. Matrices for Negatively-skewed Population: n=20

| COUNT | COLUMN TOTAL | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|-------|--------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| 7     | 51           |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |
| 8     | 61           |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |
| 9     | 71           |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |    |
|       | TOTAL        | 2| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  | 2  |

**Source:** The table and text are related to statistical analysis, likely for a specific academic or research context, but the exact nature of the analysis is not clear from the image alone.
APPENDIX F

PROGRAM FOR COMPUTING INDICES BY GAUSS
Proc to generate samples, compute and write all stats to a file. The four arguments are:

- **x** - matrix containing test scores
- **n** - number of samples desired (1000)
- **ni** - number of items in the test
- **w** - after computing **w** rows of statistics, the proc writes intermediate results to disk to save memory

```plaintext
proc (0)=boot4(x,n,ni,w);
local i, d, sub, s, stat, c, ni, nx, nv, ss, sorted, sdata,
    hunkappa, skoviac, living, brenkane, meanstd, all, y, fname,
    fout, fin, inter, b, w;

nx = rows(x);
nv = cols(x);

fname="c:\outdatasetname";
create fout = ~fname with var,6,4;
open fout = ~fname for append;

ss=50;
do while ss.<=150;
c=12;
do while c<=16;

i=1;
do until i .> n;
   if (i-1) % w == 0;
      /* Selects random sample */
         sorted=sortc((ceil(rndu(nx,1)*nx)^x),1);
         sdata=sortrad[1:ss,2:nv+1];
         "I am now working on sample "; i;
         "SS is "; ss; " with C of "; c;
      /* Computes statistics for sample, stores in row vector */
         all=(huynh(sdata[..,1],c,ni)^subkov(sdata[..,1],c,ni)^)
            (livstn(sdata[..,1],c,ni)^)~
            (swami(sdata,c,ss));
   else;
      /* Selects random sample */
         sorted=sortc((ceil(rndu(nx,1)*nx)^x),1);
         sdata=sortrad[1:ss,2:nv+1];
         "I am now working on sample "; i;
```
/* Proc to compute the Livingston k^2. */

proc lvstn(x,c,n);
local n, m, kr21, v, l, x;

m=meanc(x);

kr21= (n./(n-1)).*(1-(m.*(n-m))./(n.*stdc(x).^2));

v=stdc(x).^2;

l=((kr21.*v)+(m-c).^2)./(v+(m-c).^2);

retp(l);
endp;
/* Proc for computing Brennan and Kane's M for all variables in a matrix. The three arguments are:

  x - matrix containing test scores
  c - cutoff score (mastery criterion score)
  ni - number of items in the test

------------------------------------------*/

proc bk(x,c,ni);
  local ni, c, vx, mx, m;
  vx = (stdc(x) . "x" . "x") ;
  mx = meanc(x);
  m = 1 - (1/(ni-1)) . * ( ((mx.*(ni-mx)) - vx) ./ ((mx-c)."^2+vx) ) ;
  retp(m);
endp;
/* Proc to compute Huynh's kappa coefficient. */

The three arguments are:

- $x$ - matrix containing test scores
- $c$ - cutoff score (mastery criterion score)
- $ni$ - number of items in the test

The proc returns a row vector containing the simple normal approximation of $k$, horizontally concatenated with the simple normal approximation of $p$.

```plaintext
proc huynh(x,c,ni);
local s, ni, m, kr21, z, p1, p2, k, app;

m=meanc(x);
s=stdc(x);
z=(c-.5-m)./s;

kr21=(ni./(ni-1))*(1-(m.*(ni-m))./(ni*s.^2));

p1=cdfn(z);

p2=cdfbvn(z,z,kr21);

k=(p2-p1.^2)/(p1-p1.^2);

app=1+2*(p2-p1);

retp(k,'~app');
endp;
```
/* This procedure generates the Subkoviak F for all variables of an input matrix x, number of items ni, and a cutoff point c. */

proc subkov(x,c,ni);
local n, m, k, cuti, loop, ni, kr21, psubi, comb, term, pisum, pc, p;

m=meanc(x);
n=rows(x);

kr21=(ni/(ni-1)) * (1- (m.*(ni-m) ./ (ni*stdc(x).^2) ) ) ;

psubi=(kr21'*(.x./ni))+(((1-kr21).*(m./ni))');

cuti=c;
loop=1;
do while cuti < ni;

if loop eq 1;

/* Computes combinations */
comb=(prodc(seqa(cuti+1,1,ni-cuti)))/(ni-cuti)!;

/* Computes one term of P( X(i) >= C ) */
term=comb*(psubi.^cuti).*(1-psubi).^(ni-cuti);

pisum=term;
cuti=cuti+1;
loop=loop+1;

else;

comb=(prodc(seqa(cuti+1,1,ni-cuti)))/(ni-cuti)!

term=comb*(psubi.^cuti).*(1-psubi).^(ni-cuti);

/* Sums from X(i)=C to n */
pisum=pisum+term;
cuti=cuti+1;
endif;
endo;

p=pisum.^2 + (1-pisum).^2;

pc=meanc(pisum);
retp(pc);
endo;
Proc to compute the Swaminathan kappa for a dataset.

The three arguments are:

- \( x \) - matrix containing test scores
- \( c \) - cutoff score (mastery criterion score)
- \( ss \) - number of observations in the dataset

```
proc swami(x,c,ss);
local pc, pc, k, agree, ss, c, dum1, dum2, dumsum1, dumsum2;
/* Computes column vectors representing mastery or non-mastery on each test for each observation. From comparison of these columns, determination is made regarding the mastery-mastery, mastery-non-mastery, etc. conditions. The agree vector contains ones for the mastery-mastery and non-mastery-non-mastery conditions. */
dum1=dummydn(x[:,1],c=.5,1);
dum2=dummydn(x[:,2],c=.5,1);
agree=dum1 .eqv dum2;

pc=sumc(agree)./ss;
dumsum1=sumc(dum1);
dumsum2=sumc(dum2);

pc=((dumsum1./ss).*(dumsum2./ss))+
    (((ss-dumsum1)./ss).*(ss-dumsum2)./ss));

k=(pc-pc)/(1-pc);
retp(k);
endp;
```
APPENDIX G.
1. Algebraic Equivalence of $k^2(X, T_1)$ and $M(C)$

$$\frac{\alpha_{21}(q^2) + (\bar{x} - \bar{c})^2}{\sigma^2 + (\bar{x} - \bar{c})^2} = 1 - \frac{1}{n - 1} \left[ \frac{\bar{x}(n - \bar{x}) - \sigma^2}{(\bar{x} - \bar{c})^2 + \sigma^2} \right]$$

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

$$= \frac{n \sigma^2 - (\bar{x}(n - \bar{x}))}{n - 1} \frac{q^2 + (\bar{x} - \bar{c})^2}{\sigma^2 + (\bar{x} - \bar{c})^2}$$

$$= \frac{(\bar{x} - \bar{c})^2}{q^2 + (\bar{x} - \bar{c})^2} + \frac{-[\bar{x}(n - \bar{x}) - n \sigma^2]}{(n - 1) \left[ \sigma^2 + (\bar{x} - \bar{c})^2 \right]}$$

$$= \frac{(\bar{x} - \bar{c})^2}{q^2 + (\bar{x} - \bar{c})^2} - \frac{1}{n - 1} \left[ \frac{\bar{x}(n - \bar{x}) - \sigma^2}{\sigma^2 + (\bar{x} - \bar{c})^2} \right]$$

$$= \frac{\sigma^2 + (\bar{x} - \bar{c})^2}{\sigma^2 + (\bar{x} - \bar{c})^2} - \frac{1}{n - 1} \left[ \frac{\bar{x}(n - \bar{x}) - \sigma^2}{\sigma^2 + (\bar{x} - \bar{c})^2} \right]$$

$$= 1 - \frac{1}{n - 1} \left[ \frac{\bar{x}(n - \bar{x}) - \sigma^2}{\sigma^2 + (\bar{x} - \bar{c})^2} \right]$$

which is $M(C)$
BIBLIOGRAPHY


Huynh, H., & Saunders, J.C. (1979, April). Bayesian and empirical bayes approaches to setting passing scores on mastery tests. (Research Memorandum 79-2). Columbia, S.C: University of South Carolina.


