INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
A COMPARISON OF METHODS FOR
ESTIMATING INTERNAL CONSISTENCY RELIABILITY
OF TESTS WITH DICHOTOMOUSLY-SCORED ITEMS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
EDUCATIONAL PSYCHOLOGY
MAY 1995

By
Alfred E. Lupien

Dissertation Committee:
Harold I. Ayabe, Chairman
Shuqiang Zhang
Elizabeth Brandt
Patricia Nishimoto
Martha E. Crosby
© Copyright 1995

by

Alfred E. Lupien
Acknowledgements

There is nothing more precious in life than people. I would like to acknowledge some of individuals who have helped me fulfill this doctoral dream.

I received direct support from the members of my dissertation committee: Dr. Ayabe who peaked my interest in psychometrics and has provided on-going guidance; Dr. Zhang, my first University of Hawaii professor and one of the finest classroom teachers I have ever met; and Drs. Brandt, Crosby, and Nishimoto who trusted me to "get the job done."

No one has influenced my professional development more than Dr. Mary Colette Smith from the University of Alabama at Birmingham. I'm glad I had the courage to ask her to chair my thesis committee.

My best friend, Donna, has had the difficult task of providing day-to-day support from 2000 miles away. Thanks for listening - a lot.

Most importantly, thanks to my parents, who have been the strongest influence in my life. They have taught me more about love and life than all the books in the world.
Abstract

The use of Kuder and Richardson’s formula 20 (KR20) (1937) to estimate reliability has been controversial. The purpose of this investigation was to review internal consistency reliability estimation techniques for unidimensional tests with dichotomously-scored items. Eleven methods were compared using a series of 98 simulated item-by-person response patterns with positive off-diagonal covariances, including patterns known to reflect perfect reliability by Loevinger’s index of homogeneity (1947) and KR20. The upper limit of 1.0 was achieved in both perfect patterns only using methods described by Cliff (1984), Horst (1953), Loevinger, and Raju (1982). Lower limits of reliability were projected through linear regression. The ratio of off-diagonal covariance to test variance was used as the independent variable. Zero was included in the 95% confidence interval for Y-intercepts with Cliff’s, Horst’s, and Kuder-Richardson’s techniques. Negative Y-intercepts were computed for the techniques of Cliff, Huck (1978), Loevinger, and Winer (1971). Positive Y-intercepts were computed for the techniques of Ayabe (1994), Guttman (L₁ and L₂) (1945), Raju, and ten Berge and Zegers (1978). Between the upper and lower limits, reliability estimates generally increased as the ratio of off-diagonal covariance to total variance increased. It was concluded that the majority of estimation techniques do not meet minimum criteria for
interpretation. Only the methods of Cliff, Horst, and Raju generally met the requirements for reliability estimation techniques. Compared to KR20, the mean increases in reliability estimated with these three methods were .12 with Raju's ratio of actual to maximal KR20, .04 with Horst's method, and .00 with Cliff's $\gamma$-reliability technique.
# Table of Contents

Acknowledgements ................................................. iv
Abstract ........................................................................ v
List of Tables ............................................................ x
List of Figures ............................................................. xii
List of Abbreviations ..................................................... xiii

**Chapter 1: Introduction** ........................................... 1
   Classical Measurement Theory and Reliability ................. 3
   Definitions of Test Reliability .................................... 5
   Factors Affecting Test Reliability .................................. 9
   Internal Consistency and Homogeneity ........................... 11
   Significance of the Study ........................................... 13

**Chapter 2: Review of Literature** ............................... 15
   The Development of Kuder-Richardson Formula 20 .......... 15
   Support for the Kuder-Richardson Model ....................... 24
   Comparing KR20 to Split-Half Reliability ..................... 24
   Extensions of the Kuder-Richardson Model ................. 27
   Alternatives to the Kuder-Richardson Model .................. 31
      The Phi Coefficient ........................................... 32
      A Phi Alternative ............................................. 36
      Guttman’s Lower Bounds ..................................... 41
      The Greatest Lower Bound ................................... 45
      An Algorithm for Estimating the Maximum Split-Half Homogeneity: A Related Concept ..................... 50
      Criticisms of Loevinger and Horst ........................... 58
      The Ratio of Actual to Maximal KR20 Estimates .......... 61
Guttman Data Matrices with Column Substitution .... 132
Kuder-Richardson Data Matrices with Column Substitution .......................... 132
Random Data Matrices ............................................................................. 135
Chapter 5: Discussion .............................................................................. 138
Reiteration of Findings ........................................................................... 139
The Upper Limit of Reliability ................................................................. 139
The Lower Limit of Reliability ................................................................. 139
Goodness of Fit ......................................................................................... 140
Association Among Estimation Techniques ........................................... 141
Practical Interpretation of Findings ......................................................... 141
Relation of Findings to Research Design ................................................ 143
Relation of Findings to Review of Literature .......................................... 144
Limitations of Phi .................................................................................... 144
Limits of Reliability .................................................................................. 147
General Criticisms ................................................................................... 148
Standards for Comparison of Estimation Methods ............................... 149
Conclusions ............................................................................................. 150
Significance of the Investigation ............................................................. 150
Implications for Future Research ........................................................... 153
References ............................................................................................... 154
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Summary of Proposed Techniques for the Estimation of Internal Consistency Reliability</td>
<td>74</td>
</tr>
<tr>
<td>3.</td>
<td>Frequency Distribution of Row and Column Numbers Selected for Substitution</td>
<td>93</td>
</tr>
<tr>
<td>4.</td>
<td>Frequency Distribution of Mean Item Variances by Item-by-Person Matrix</td>
<td>95</td>
</tr>
<tr>
<td>5.</td>
<td>Frequency Distribution of Standard Deviation by Item-by-Person Matrix</td>
<td>95</td>
</tr>
<tr>
<td>6.</td>
<td>Frequency Distribution of Kuder-Richardson Reliability Estimates by Item-by-Person Matrix</td>
<td>96</td>
</tr>
<tr>
<td>7.</td>
<td>Overall Sample Characteristics</td>
<td>106</td>
</tr>
<tr>
<td>8.</td>
<td>Sample Characteristics for Guttman Scales with Row Substitutions</td>
<td>107</td>
</tr>
<tr>
<td>9.</td>
<td>Sample Characteristics for Guttman Scales with Column Substitutions</td>
<td>109</td>
</tr>
<tr>
<td>10.</td>
<td>Sample Characteristics for KR Scales</td>
<td>110</td>
</tr>
<tr>
<td>11.</td>
<td>Sample Characteristics for Random Matrices</td>
<td>112</td>
</tr>
<tr>
<td>12.</td>
<td>Reliability Estimation Techniques Achieving Upper Limit of 1.0</td>
<td>113</td>
</tr>
<tr>
<td>13.</td>
<td>Estimated Lower Limits of Reliability Estimation Techniques</td>
<td>115</td>
</tr>
<tr>
<td>14.</td>
<td>$R^2$ and $F$-Ratios for Goodness of Fit and Linearity</td>
<td>116</td>
</tr>
<tr>
<td>15.</td>
<td>Product-Moment Correlation Coefficients Among Reliability Estimates for Complete Data Set</td>
<td>120</td>
</tr>
<tr>
<td>16.</td>
<td>Product-Moment Correlation Coefficients Among Reliability Estimates for Guttman Row Data Sets</td>
<td>121</td>
</tr>
<tr>
<td>17.</td>
<td>Product-Moment Correlation Coefficients Among Reliability Estimates for Guttman Column Data Sets</td>
<td>122</td>
</tr>
</tbody>
</table>
18. Product-Moment Correlation Coefficients Among Reliability Estimates for Kuder-Richardson Data Sets .................................................. 123


20. Ascending Rank-Order of Mean Reliability Coefficients .................................................. 125

21. Differences in Reliability Coefficients Between Estimation Techniques and KR20 ................. 126

22. Characteristics of Reliability Estimates from Data Matrices Excluded from Analysis ............... 127
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Ayabe (1994) Test Item-by-Person Score Matrices</td>
<td>77</td>
</tr>
<tr>
<td>2.</td>
<td>Rearranged Ayabe Test Item-by-Person Score Matrices C and D</td>
<td>81</td>
</tr>
<tr>
<td>3.</td>
<td>Perfect Item-by-Person Score Matrices</td>
<td>91</td>
</tr>
<tr>
<td>4.</td>
<td>Residual Scatterplot for $r_A$ and $L_1$</td>
<td>118</td>
</tr>
<tr>
<td>5.</td>
<td>Residual Scatterplot for $H_1$ and $L_1$</td>
<td>119</td>
</tr>
<tr>
<td>6.</td>
<td>Reliability Estimates as a Function of Off-Diagonal Covariance: Total Sample</td>
<td>129</td>
</tr>
<tr>
<td>7.</td>
<td>Reliability Estimates as a Function of Off-Diagonal Covariance: Guttman Sample with Row Substitution</td>
<td>131</td>
</tr>
<tr>
<td>8.</td>
<td>Reliability Estimates as a Function of Off-Diagonal Covariance: Guttman Sample with Column Substitution</td>
<td>133</td>
</tr>
<tr>
<td>9.</td>
<td>Reliability Estimates as a Function of Off-Diagonal Covariance: Kuder-Richardson Sample Column Substitution</td>
<td>134</td>
</tr>
<tr>
<td>10.</td>
<td>Reliability Estimates as a Function of Off-Diagonal Covariance: Random Substitution</td>
<td>136</td>
</tr>
<tr>
<td>11.</td>
<td>Comparison of Selected Reliability Estimates with KR20</td>
<td>137</td>
</tr>
</tbody>
</table>
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>Error</td>
</tr>
<tr>
<td>$H_i$</td>
<td>Loevinger's index of homogeneity</td>
</tr>
<tr>
<td>$i$</td>
<td>Item (such as the $i$th item)</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of test items</td>
</tr>
<tr>
<td>KR20</td>
<td>Kuder-Richardson formula 20</td>
</tr>
<tr>
<td>$L$</td>
<td>Guttman's estimates for reliability ($L_1 - L_n$)</td>
</tr>
<tr>
<td>$M$</td>
<td>Mean</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of subjects or test takers</td>
</tr>
<tr>
<td>$p$</td>
<td>Item Difficulty or probability of answering test item correctly</td>
</tr>
<tr>
<td>$pq$</td>
<td>Item variance</td>
</tr>
<tr>
<td>$S_i^2$</td>
<td>Variance of actual test scores</td>
</tr>
<tr>
<td>$r$</td>
<td>Product-moment correlation coefficient</td>
</tr>
<tr>
<td>$r_{ij}$</td>
<td>Covariance of $i$th and $j$th items</td>
</tr>
<tr>
<td>$r_A$</td>
<td>Ayabe's estimate of reliability</td>
</tr>
<tr>
<td>$r_H$</td>
<td>Horst's estimate of reliability</td>
</tr>
<tr>
<td>$r_{He}$</td>
<td>Huck's estimate of reliability</td>
</tr>
<tr>
<td>$r_R$</td>
<td>Raju's estimate of reliability</td>
</tr>
<tr>
<td>$r_W$</td>
<td>Winer's estimate of reliability</td>
</tr>
<tr>
<td>$S_p$</td>
<td>Standard deviation of item difficulties</td>
</tr>
<tr>
<td>$SS$</td>
<td>Sum of Squares</td>
</tr>
<tr>
<td>$T$</td>
<td>True test score</td>
</tr>
<tr>
<td>$X$</td>
<td>Actual test score</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Alpha reliability (equivalent to KR20 when items are scored dichotomously)</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Goodman-Kruskal gamma</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Lambda, Guttman's theoretical reliability coefficient</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mu, ten Berge and Zegers' reliability coefficient</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Phi coefficient</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>Sigma, summation</td>
</tr>
</tbody>
</table>
CHAPTER 1
Introduction

There is no point to testing knowledge or abilities if information obtained from the test is inaccurate. Considerable attention has been focused on verifying that a test measures the intended concept (validity), and that the measurement is consistent (reliability). Although techniques used for estimating validity and reliability are usually discussed separately, the two test characteristics are related inextricably. Validity is limited by reliability such that the maximum validity coefficient is the square root of the reliability coefficient (Gullicksen, 1987; Sax, 1989). Because of the dependent relationship of validity upon reliability, the reliability of a test has a significant effect on the efficacy of any test. According to Ebel (1972), "the reliability coefficient provides the most revealing index of quality that is most ordinarily available" (p. 407).

Reliability has been defined as the stability of a test from one use to the next (Vogt, 1993). The definition of reliability implies the administration of either two parallel forms or the same test twice. Unfortunately, it is not always possible to test students twice. It would be beneficial if test reliability information could be obtained from only one test administration.
In 1937, Kuder and Richardson introduced a series of formulas for estimating the internal consistency of a test from a single administration. Internal consistency has been defined as the extent to which test items measure the same concept (Vogt, 1993). With the ability to estimate internal consistency from a single test administration, Kuder and Richardson's formula 20 (KR20) has become one of the most widely-used methods for estimating reliability. Assumptions made in the derivation of KR20 have made the formula controversial as well. Attempts at the re-derivation of KR20 using different, less restrictive, assumptions has been described by the Educational Testing Service as "the most popular indoor sport in psychometrics" (cited in Kuder, 1991, p. 873). In 1991, Kuder called for a review of KR20 and proposed alternatives.

This dissertation responds to Kuder's request. In the nearly 60 years since KR20's introduction, alternative methods for the estimation of internal consistency reliability have been proposed in the literature. Most commonly, authors either introduce new estimation techniques with little substantiating evidence, or present comparisons of a select few estimation procedures. The purpose of this investigation is to review methods for the estimation of internal consistency reliability for unidimensional tests with dichotomously-scored items, and subject selected methods to a series of item-by-person response patterns.
The goal of the investigation is to provide educators a comprehensive analysis and comparison of available methods for the estimation of internal consistency reliability.

**Classical Measurement Theory and Reliability**

An individual’s observed test score \( X \) is divided into two components, true \( T \) and error \( E \) scores, such that an individual’s observed score is the sum of the true and error components \( X = T + E \). Only the observed score can be directly measured, but it is the true score that is of most interest in quantifying an individual’s knowledge or ability, the purpose for which the test was administered.

True and error scores have certain properties that serve as the fundamental assumptions of classical measurement theory. The true score has been defined as "the average of the scores earned by an individual on an unlimited number of perfectly parallel forms of the same test" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1985, p. 94). Ideally, the true score is the actual score the test taker should receive.

The difference between the observed and true scores is termed error. By definition, error has certain properties: (a) there is no correlation between true and error scores, (b) there is no correlation between error scores on parallel tests, and (c) the mean of error scores equals zero with
repeated measurement (Gullicksen, 1987). Because the average error becomes zero over trials, and the true score is the difference between observed score and error, then the mean of observed score becomes the best estimate of true score.

Just as the individual test score can be subdivided into true and error components, the variance of observed test scores can be subdivided such that the variance of actual test scores ($S_X^2$) equals the sum of true ($S_T^2$) and error variances ($S_E^2$), that is $S_X^2 = S_T^2 + S_E^2$. The reliability ($R$) of a test has been defined conceptually as the ratio of true variance to observed test variance ($R = S_T^2 / S_X^2$) (Gullicksen, 1987). The variance-based definition of reliability establishes a theoretical mathematical range for reliability. If there is no error, the observed and true variances are the same and the variance ratio equals 1. This unity value is considered to be the upper limit for test reliability. If all observed variance is error, then there is no true variance and the reliability coefficient becomes zero.

The extent to which a test is considered reliable is reported as a coefficient typically computed as the product moment correlation between two sets of scores on equivalent tests obtained independently from members of the same group (Ebel, 1972). At least two independent measures of the same trait are required for each group member.
Since computation of the reliability coefficient requires a second test administration, classical methods for estimating reliability do not measure a property of the test pre se, but rather, a property of the test within the context of its measurement (Hambleton, Swaminathan, & Rogers, 1991). Item characteristics are computed based on the performance of the individuals taking the test. Subjects from a different population may respond differently to the same items. Examinee scores are also dependent on the particular set of items comprising a test and overall test scores are a function of the characteristics of the group of items selected. Discontent with the classical measurement model has led to the development of alternative models for test development such as item response theory which focuses on items with characteristics assumed to be independent of the test in which they are administered and the subjects who are responding to the items.

Definitions of Test Reliability

Despite the critical nature of reliability with respect to the precision of a test, the definition of reliability remains contextual. Use of the term reliability varies among authors and complicates direct comparison of estimation techniques and interpretation of investigation results.

A general definition is contained in the Standards for Educational and Psychological Testing where reliability is
defined as "the degree to which test scores are free from errors of measurement" (AERA, APA, & NCME, 1985, p. 19). Four causes of measurement error are identified: random response variability, changes in the individual taking the test, differences in item content between test forms, and differences in the administration of the tests. Differences in test scores related to "maturation, intervention, or some other event that has made these differences meaningful or if inconsistency of responses is relevant to the construct being measured" (AERA, APA, & NCME, 1985, p. 19) were not defined as error.

Three types of reliability have been described by Sax (1989): equivalence, stability, and internal consistency. Equivalence refers to the administration of two or more parallel forms of an examination to the same people at approximately the same time. Equivalence reliability is estimated as the product moment correlation of scores from individuals taking both forms of the test. Equivalence is influenced by differences in test content between the two forms, random variation in response, and differences in how each of the tests was administered.

Stability can be determined by correlating scores from repeated administration of an identical test to the same individuals after a period of time. Reliability obtained by repeated administration is referred to as test-retest reliability. The test-retest method remains the most
popular technique to estimating the stability of a test over time; but the method is vulnerable to artifact from random variability in response, changes in the individuals taking the test, and differences in the method by which the tests were administered.

The third method of reliability is internal consistency. While stability and equivalence require two test administrations, internal consistency reliability is a type of reliability estimate based on the administration of a single test form (AERA, APA, & NCME, 1985). With only one test administration, the internal consistency coefficient is not influenced by changes in the individuals taking the test, differences between test forms, or differences between test administrations. Of the four general factors influencing test reliability identified by the AERA, APA, and NCME, only random response variability is operative.

One of the earliest and well-known techniques for the estimation of internal consistency is the split-half method where the items from a single test are divided into two separate halves, either arbitrarily or using some criteria such as item content, discrimination, or difficulty. The individual’s scores from the two test halves are compared. The obtained product moment correlation coefficient is adjusted to correct for differences based on test length using the Spearman-Brown Prophecy Formula (Sax, 1989).

According to Ebel (1972), the split-half method is probably
one of the best techniques for estimating reliability in circumstances where test answers will be corrected for guessing, or where item weighting will be used.

Since there are \( K! / \{2[(K/2)!]^2 \} \) methods of splitting a test in half, where \( K \) = the number of test items, the actual value of the computed reliability estimate fluctuates based on which split is computed. Rather than using an arbitrary splitting, the expected distribution of reliability coefficients for a given test could be computed. Once the distribution was known, the most representative value, such as the mean, could be used as the best estimate of reliability (Cronbach, 1951; Kuder & Richardson, 1937).

Computing all possible split-half coefficients could be laborious for tests with large numbers of items. For example, a test with only 10 items would have 126 possible splits. A 20-item test would have 92,378 splits.

In 1937, Kuder and Richardson developed a series of equations for estimating internal consistency reliability by calculating the correlation coefficient between two tests, one real and one hypothetical. Cronbach (1951) demonstrated that his coefficient alpha (\( \alpha \)), which was mathematically equivalent to KR20 when test items are scored dichotomously, was equal to the mean of all possible splits of a test into halves. He speculated since \( \alpha \) was mean of all possible split-halves, about half of the coefficients would be greater than \( \alpha \) and the other half would be lower assuming
that split-half coefficients are distributed symmetrically. [The distribution of split-half coefficients was found later to be asymmetrical by Heim and Ayabe (1986)]. Cronbach also proposed that $\alpha$ was the exact parallel-form correlation when the mean covariance between parallel items was equal to the mean covariance between unpaired items. According to Cronbach (1951), this was the least restrictive assumption for "proving" the Kuder-Richardson formula.

Cronbach's interpretation of Kuder and Richardson's work suggests why KR20 has become a popular formula for estimating reliability. The comparison of the real test to its hypothetical equivalent represents a "zero time interval" test-retest to estimate stability unaffected by random variation in response. By accepting Cronbach's proposal that KR20 was a version of parallel form reliability, equivalence is approximated without regard for differences in test content between forms. Reliability estimation with one test administration also minimizes changes in the examinees and testing conditions.

Factors Affecting Test Reliability

Symonds (1967) described 22 factors affecting the reliability of a test. Six factors are considerations in test construction and relate to this investigation:

1. Number of test items. Reliability increases as the number of items increases. The influence of the number
of test items is described by the Spearman-Brown prophecy formula.

2. The range of item difficulty. The narrower the range of item difficulty, the greater the reliability. Items that are answered correctly (or incorrectly) by all individuals do not contribute to variability within the test and decrease the number of functional test items.

3. Evenness in scaling. The result of developing tests with items at the same level of difficulty is equivalent to reducing the number of test items. All items of equal difficulty should be answered either correctly or incorrectly. The extreme case is two sets of items: those answered correctly by all students, and those answered incorrectly by all students. This situation reduces the test to two items. Optimally, the test will be evenly scaled across a range of item difficulties.

4. Interdependence of test items. Lower estimates of reliability will be achieved if the answer to one item is suggested by another item, or if the meaning of one item is dependent upon a previous item.

5. Guessing. Reliability decreases as the likelihood of guessing the correct answer increases.

6. Homogeneity. If items test different concepts, then reliability will decrease.
Internal Consistency and Homogeneity

Internal consistency and homogeneity are closely related concepts. Internal consistency should be used to describe the degree of inter-relatedness among items regardless of whether the items are measurable in a single dimension. The most common scenario is where the majority of item variance is accounted for by a first general factor; however, high internal consistency reliability could also be computed in situations where the items are affected by several common factors (Green, Lissitz, & Mulaik, 1977; Hattie, 1984; White & Saltz, 1957).

Homogeneity is limited to describing a set of items measuring a single common factor in their true score component. In a perfectly homogenous test there may be no instance where a subject passes an item after failing an easier one (Loevinger, 1947). The purest example of a homogeneous test is the Guttman scalogram. Items are related to each other by means of an increasing monotonic function. The entire item-by-person score matrix could be reproduced from limited information: the total number of correct items and the item difficulties (White and Saltz, 1957). The monotonic relationship between items has been termed tau-equivalence by Novick and Lewis (1967).

According to Loevinger (1947), the ability to construct a definitive arrangement of test items provides a criterion for estimating homogeneity, rather than the ambiguous
criteria used for internal consistency. High homogeneity implies high internal consistency, but high internal consistency does not necessarily mandate high homogeneity. Not all single-factor tests are highly homogeneous.

Whether a model allowed for error was a second characteristic used to distinguish between homogeneity and reliability. Low homogeneity implied item heterogeneity whereas low reliability indicated error variance (White & Saltz, 1957).

White and Saltz (1957) argued that homogeneity requires stringent conditions: (a) factor(s) determining subject responses did not change during the test, (b) factor(s) determining subject responses were the same for all students, and (c) all items were identical in the factors determining the responses they elicit. The same conditions were true for reliability as well, but they are seldom met, if ever. According to White and Saltz (1957), the inability to meet these requirements are not a fault of Guttman and his scalogram, but point to the complexity of human beings. At the same time,

[T]here is noting wrong in continuing to assume that many human abilities, attitudes, and traits are unidimensional continua, but we should be fully aware that this is at best a useful first approximation, and that an appreciable proportion of the information in our raw data will thereby be
sacrificed at the altar of error variance. (p. 98)

Significance of the Study

Reliability is an important concept at all levels of educational testing from school classrooms to post-graduate professional certification. A recent issue of Evaluation and the Health Professions (La Duca, 1994) explored the validity of examinations used for licensure and certification of health care providers. Given the mathematical relationship between validity and reliability, questions about the reliability could have been asked as well. Designers of large-scale examinations should be able to demonstrate test reliability easily with their expertise in psychometrics, large pools of test questions with established item characteristics, and access to sophisticated computer hardware and software.

Most testing probably occurs in classrooms at the elementary, secondary, and college levels. At a major university as many as 10,000 tests may be administered annually and thousands of grades are assigned (D. Roberts, personal communication, October 4, 1994). The individuals who design and administer the majority of tests, classroom teachers, may not be as fortunate as designers of large-scale tests. Even if the resources are available, teachers may lack the measurement expertise or computer skills required for sophisticated test analysis.
Testing for national certification and classroom performance are examples of separate but inter-related situations. Just as total test score has been defined as the sum of item scores, a student's opportunity to take a national certification examination is affected by the series of grades received in previous course work. Individuals who are being measured by tests at every level are entitled to fair and equitable testing.

There is a need to identify reliability estimation techniques that are simple to compute and require only limited resources. Despite the recent trend toward item response theory and away from classical measurement theory, classical techniques may be useful for the classroom teacher. They are easy to compute and the criticisms of classical measurement theory may be less applicable in classroom settings since generalization of test scores to differing populations is not a primary concern in classroom test design.

Techniques for the estimation of test reliability that are based in classical measurement theory remain viable options for the classroom teacher. This investigation has been conducted with the intention responding to Kuder's request for a review KR20 and alternatives. The perspective taken throughout the investigation is that of the classroom teacher, the individual most likely to use classical methods for the estimation of reliability.
CHAPTER 2

Review of Literature

Unsatisfied with the models available for the estimation of reliability, Kuder and Richardson (1937) introduced a technique for the estimation of reliability which required only one test administration. Kuder and Richardson observed that the reliability estimates obtained using previous methods yielded unpredictable results between them. Knowing the reliability coefficient estimated by one of the methods did not make it possible to predict the reliability estimate calculated by the other methods.

Kuder and Richardson proposed a method for estimating test reliability by computing the correlation between the actual test and its hypothetical equivalent. They believed their estimate would be applicable to any unidimensional test where items were unit-weighted. Inter-item coefficients were allowed to vary between their possible limits, as were varying proportions of correct answers. Items did not need to be equally difficult or equally correlated with other items.

The Development of Kuder-Richardson Formula 20

The Kuder-Richardson method was an extension of the parallel forms method for estimating reliability. Instead of using two test forms, the actual test was compared with its hypothetical equivalent. Because reliability had been defined as the correlation between two forms of a test, the
The coefficient of reliability could be estimated by computing the correlation between the actual test and its hypothetical counterpart.

When calculating the product moment correlation coefficient, the denominator of the correlation equation consisted of the sum of the item variances ($s_i^2$) plus twice the sum of inter-item covariances ($r_{ij}s_is_j$, $i \neq j$, where $i$ and $j$ represent test items). The sum of covariances ($s_i^2 + 2r_{ij}s_is_j$, $i \neq j$) was equal to the total variance of the test.

In the numerator, the inter-item covariances were the same as in the denominator ($r_{ij}s_is_j$, $i \neq j$). Item variances; however, needed to be multiplied by the reliabilities of the parallel items ($r_u$). Kuder and Richardson's conceptual formula for reliability, KR2, was

$$r_{tt} = \frac{\sum r_{ij}s_i^2 + 2\sum r_{ij}s_is_j}{\sum s_i^2 + 2\sum r_{ij}s_is_j}, \text{ where } i \neq j.$$  

To compute the coefficient of reliability, values for three components were required: item variances ($s_i^2$), inter-item correlations ($r_{ij}$, $i \neq j$), and the correlation of the item with its hypothetical equivalent ($r_u$). Only $s_i^2$ and $r_u$ could be calculated directly. The difficulty in estimating test reliability ($r_a$) was in the estimation of item reliability.
If $r_n$ could be computed, the reliability of the test would be known.

Since the denominator of Kuder and Richardson's (1937) equation was equal to the observed test variance ($S_x^2$), this value could be used as the denominator of the equation. The portion of the numerator in common with the denominator, $2r_{ij}s_is_j (i \neq j)$, could be replaced with $S_x^2 - \Sigma pq$, where $\Sigma pq$ represented the item variances ($s_i^2$). Reliability was expressed in Kuder and Richardson's formula 3 as

$$r_{tt} = \frac{\sum_{i=1}^{K} r_{ii}pq + (S_x^2 - \sum_{i=1}^{K} pq)}{S_x^2}$$

where $K = \text{number of items}$

$p = \text{probability of answering item correctly}$

$q = (1 - p)$

$S_x^2 = \text{test variance}$. Compared to KR2, the KR3 equation simplified calculation; however, the problem of estimating item reliabilities remained.

As long as tests were assumed to be unidimensional, it could be anticipated that each item and the test would measure the same concept. If this were true, then
where $r_i$ was the correlation of the individual item to the overall test score. This equation could be manipulated and solved for $r_i$ such that

$$r_{ii} = \frac{r_{it}^2}{r_{tt}}.$$ 

Substituting this approximation of $r_i$ into KR3, and solving for $r_i$, resulted in formula 8 (KR8):

$$r_{tt} = \frac{S_x^2 - \Sigma pq}{2S_x^2} \pm \sqrt{\frac{\Sigma r_{it}pq}{S_x^2} + \left(\frac{S_x^2 - \Sigma pq}{2S_x^2}\right)^2}.$$ 

Kuder and Richardson used only the positive solution of the radical. With the KR8 formula, test reliability could be calculated from the number of items on test, the difficulties of items ($p$), item-test correlations ($r_i$), and the variance of the test total scores ($S_x^2$). The only assumption made was that the matrix of inter-item correlations had rank of one.
In circumstances where it was permissible to assume that the difficulty of items was allowed to vary over a wide range, the correlation between items was constant and equal to the average correlation between items, then the reliability coefficient may be estimated as (KR9)

\[ r_{tt} = \overline{r}_{ii} \frac{(\Sigma \sqrt{P_i Q_i})^2}{S_x^2} . \]

Since the difference between the numerator and denominator of the general reliability equation was the substitution of \( \Sigma r_{ii} p q \) in the numerator for \( \Sigma p q \) in the denominator, total test variance may be re-written as

\[ S_x^2 = (\Sigma \sqrt{pq})^2 \overline{r}_{ii} - \Sigma \overline{r} p q_{ii} + \Sigma p q \]

and then solved for \( r_{ii} \) so that

\[ \overline{r}_{ii} = \frac{S_x^2 - \Sigma p q}{(\Sigma \sqrt{pq})^2 - \Sigma p q} . \]

Finally, KR14 emerged when the average \( r_{ii} \) was substituted into KR9, such that
Kuder and Richardson (1937) recommended KR14 for situations where inter-item correlations were approximately equal.

If item standard deviations were assumed to be equal, then $r_u$ could be re-written as (KR15)

$$r_{tt} = \frac{I_{ii}K^2p\overline{q}}{S_x^2}.$$ 

Following the logic of their previous derivations, test variance was re-written as

$$S_x^2 = K\overline{pq}[1+(K-1)r_{ii}] .$$

Solving the equation for $r_{ii}$, substituting this value into the previous formula (KR15), and simplifying resulted in formula 20 (KR20) for the estimation of reliability:

$$r_{tt} = \frac{K}{K-1} \cdot \frac{S_x^2-K\overline{pq}}{S_x^2} .$$
Although requiring the assumption of equal item covariances, KR20 required less computation than Kuder and Richardson's previous, more exact formulas. Because KR20 required approximation of average item variance, it may not seem as useful as the more precise KR14; however, reliability estimates obtained with KR20 did not vary by more than .001 from KR14 (Kuder & Richardson, 1937).

If all items were of the same difficulty level, then KR21 could be used to estimate reliability, such that

\[ r_{tt} = \frac{K}{K-1} \cdot \frac{S_x^2 - K \overline{p} q}{S_x^2}, \]

where the average value of \( p \) was

\[ \overline{p} = \frac{\Sigma X}{KN} = \frac{M_x}{K} \]

where

\( \Sigma X = \) the sum of test scores

\( N = \) the number of subjects

\( M_x = \) the mean of test scores.

Kuder and Richardson emphasized that KR20 required item variances, whereas KR21 required item difficulties. If items were equally difficult, then the two values would be
the same. Otherwise, the reliability estimate from KR20 would exceed that from KR21.

A criticism of the Kuder and Richardson method was the potential to calculate negative reliability estimates. Kuder and Richardson believed the reliability of a test was a function of the extent to which test items were positively intercorrelated. Reliability estimates were positive when test variance \((r_n)\) exceeded item variance and \(r_n\) was positive for any positive average inter-item correlation. Negative reliability was inadmissible.

Kuder and Richardson (1937) provided data on five college achievement tests. Reliability coefficients were computed using KR21 and Spearman-Brown corrected split-half techniques. All reliability estimates exceeded .863. In each of the five cases, the KR21 estimate was lower than the split-half method. The difference between the two estimates ranged between .009 and .024.

Kuder and Richardson also presented reliability estimates calculated by four of their formulas for these tests. The highest estimates were computed using KR8. Formulas 14 and 20 yielded results similar to each other, but less than KR8. The lowest estimates were calculated using KR21. The largest difference between KR8 and KR21 was .090.

Additional data were provided by Richardson and Kuder in 1939. Comparing their estimates to the "true"
reliability of a test, all Kuder-Richardson formulas underestimated the actual coefficient. Richardson and Kuder (1939) believed it was better to overestimate measurement error than to underestimate it. Specifically regarding the split-half method, a significant drawback was that no one can be sure of the direction by which the estimate errs.

The authors provided two additional examples of the reliability estimates using their formulas. Test A consisted of 90 items. Kuder and Richardson computed reliability estimates using split-half, KR8, KR14, KR20, and KR21. Kuder-Richardson coefficients were less than the estimate computed using the split-half technique. Results from the sample tests were consistent with the authors' hypothesized order of theoretical exactness: KR8, KR14, KR20, KR21. The greatest difference was .05, between the split-half and KR21 methods. Differences among Kuder-Richardson methods were .03 or less. Test B was described as a large scale reading test. No split half information was reported. The difference between KR8 and KR21 was .02. In contrast with their original article where Kuder and Richardson (1937) recommended KR21, in 1939 the investigators advocated using KR20. They concluded that KR20 was adequate in most situations, producing an estimate identical to, or slightly lower than more rigorous KR8.
Support for the Kuder-Richardson Model

Kuder and Richardson's formula 20 has been re-derived by other investigators. Cronbach (1951), Dressel (1940), Gulliksen (1987), Hoyt (1941), and Jackson and Ferguson (1941) re-derived KR20 using the less restrictive assumption that the average covariance among "corresponding items" was equal to average among non-corresponding items.

Comparing KR20 to Split-Half Reliability

At the time of Kuder and Richardson's introduction of KR20, the prevailing method for estimation of reliability was the split-half method. Kelly (1942) advocated improving split-half reliability by making the split tests as similar as possible through matching of item content and difficulty. Cronbach (1951) also believe that random splits would yield coefficients lower to a parallel form (or planned split). He computed test reliability coefficients from a 68-item test of mechanical reasoning and a 10-item morality scale, each administered to 250 boys. For the 68-item test, he deleted 8 items that were omitted by some students. For each test, he selected 50 papers on which to conduct an item analysis. Following the item analysis, he used the remaining 200 subjects and split each test randomly, with each half having items of comparable difficulty (type I splits), and with each split based on both difficulty and apparent content (type II split). He also calculated his proposed coefficient $\alpha$, which is equivalent to KR20 when
test items are scored dichotomously. For the 60 items of the mechanical reasoning test, $\alpha$ was calculated as .811. Fifteen random splits were conducted. Reliability coefficients ranged from .779 to .860, with a mean of .810. Ten type I splits resulted in a range of reliability estimates from .798 to .846, with a mean of .820. Eight type II splits yielded a range of estimates from .801 to .833 and a mean of .817.

For the 10 items of the morality scale, $\alpha$ was calculated to be .715. All 126 possible random splits were computed with the range of reliability coefficients from .609 to .797 and a mean of .715. Six type II splits yielded reliability estimates from .681 to .780, with a mean of .737. Cronbach concluded that knowing content and difficulty did not permit the test designer to make a comparable half-test resulting in an estimate better than the random split. None of the splits resulted in a coefficient remarkably different from $\alpha$ in either test. Variation from sample to sample would probably contribute more variation than that from the use of $\alpha$ or from any of the splits (Cronbach, 1951).

Unable to support Kelly's proposal that homogeneity between halves would yield higher reliability estimates, Cronbach (1951) studied the problem further through factor analysis. Even tests appearing heterogeneous were frequently saturated with the first factor among items.
Cronbach’s finding led him to examine certain theories of test design. He noted that for a test to be interpretable, it was essential that all items be factorially similar, otherwise it is difficult to ascribe meaning to the score from that test.

According to Cronbach (1951), $\alpha$ estimated the proportion of test variance due to all common factors among the items. Unless the inter-item correlation matrix had a rank higher than 1, $\alpha$ underestimated common-factor variance, but not seriously unless the test contained distinct clusters. The proportion of test variance due to the first factor was the essential determiner of the interpretability of scores. Cronbach’s $\alpha$ was an upper-bound for this.

Cronbach also observed one limitation of $\alpha$, and many other methods for estimating reliability. Ideally, the internal consistency of a test should be independent of test length. Unfortunately $\alpha$, as well as the Loevinger-Ferguson indices, increased as a function of test length.

The usefulness of KR20 was explored also by Lord (1955). Lord intended to determine the reliability coefficient resulting from new definitions of parallelism. The coefficient for "matched-form" reliability leads to a least upper bound for test reliability attained by careful matching of items in parallel form test. He developed a technique for computing a least upper bound through precise matching of items. He acknowledged the technique had only
theoretical significance since it required knowing the curvilinear multiple correlation of responses to each item with various common factors measured by the same items. It could be used; however, to assess the reasonableness of Kuder-Richardson reliability estimates. Using an algebra test where items were matched on common factors so that the upper bound parameter was estimated, Lord compared reliability estimates. For the entire 28-item test, the KR20 estimate of reliability was .805 compared to .814 by the matched form. For the 10 most valid items, estimates were again similar. The estimate of reliability by KR20 was .745 compared to .757 by the matched-form. Lord concluded that in most cases KR20 would approach the least upper bound of test reliability, as computed by the matched-forms coefficient.

Extensions of the Kuder-Richardson Model

Hoyt (1941) used analysis of variance (ANOVA) to achieve the same reliability estimate as computed using KR20. It was assumed that for each individual, the true scores, $\pi_{ij}$, remained constant across all items, and the error scores, $\epsilon_{ij}$, were normally distributed and uncorrelated across items. Hoyt used a 2-way repeated measures ANOVA with 1 replicate per cell to estimate the reliability of a test following the model: $X_{ij} = \rho + \pi_i + \rho_j + \epsilon_{ij}$, where $\rho$ was a constant reflecting the average ability level of the group of individuals with respect to the particular set of
questions, \( \pi_i \) was the relative ability of the \( i \)th examinee, 
\( \rho_j \) was an index of the comparative difficulty of the \( j \)th item, and \( \epsilon_j \) was an error of measurement. Hoyt’s computational formula was

\[
I = \frac{MS_{\text{exam}} - MS_{\text{res}}}{MS_{\text{exam}}}. 
\]

Hoyt’s method yielded an estimate of the percentage of variance regarded as not related to the instrument. Extended evaluation of sum of squares for items (\( SS_{\text{item}} \)) would allow examination of heterogeneity of item difficulties and examination of the sum of squares for examinees (\( SS_{\text{exam}} \)) would provide additional information about individual differences among students (Hoyt, 1941).

Hoyt provided no evidence for the veracity of his reliability estimate. His interest was in the presentation of an alternative method for the computation of KR20. An advantage of Hoyt’s technique was that it could also be used in circumstances where differential weighting of items was necessary or desirable (Kaitz, 1945).

The use of ANOVA to estimate reliability was extended by Winer (1971) and Huck (1978). Winer (1971) modified Hoyt’s formula to eliminate the biased estimate of the ratio of true score variance to error variance. The correction
included \( m' \) where

\[
m' = \frac{(N-1)(K-1)}{(N-1)(K-1)-2}.
\]

Winer's modification resulted in the following formula:

\[
r = \frac{MS_{exam} - m' MS_{res}}{MS_{exam}}.
\]

Huck cautioned that Winer's modification provided an accurate estimate if and only if the assumption of a constant true score \( (\pi_{ij}) \) was valid. Neither Hoyt (1941) or Winer (1971) accounted for a \( \pi\rho_{ij} \) interaction. In fact, Winer believed the ANOVA model could not be used to estimate reliability when true scores changed irregularly from one measurement to the next.

According to Huck (1978), when it was reasonable to assume that \( \pi_{ij} \) varied across items, then the psychometric model for fallible scores on specific items should remain

\[
X_{ij} = \pi_{ij} + \epsilon_{ij},
\]

but the score underlying ANOVA should incorporate an interaction component, such that
$X_{ij} = \rho + \pi_i + \rho_j + \pi_{ij} + \epsilon_{ij}$

where $\pi_{ij}$ represented the interaction component reflecting "the realistic possibility that the reason why two examinees have similar observed scores on one item but quite different scores on a second item is not likely to be related to measurement error" (Huck, 1978, p. 730). Comparing the models of Winer and Huck, $\rho$, $\pi_i$, and $\rho_j$ are the same, but in Huck's model $\epsilon_{ij}$ were subdivided into $\pi_{ij}$ and $\epsilon_{ij}$. In the Hoyt model, if an individual did better than expected on an item, based on knowledge of the examinee's total score and the group's performance on that item, then it is concluded that the result must be measurement error. In Huck's model, something extraneous may have contributed to the unusually high score. Revising the error term did not affect the theoretical model (Huck, 1978).

Although there had been suggestions by Lu (1971) and Rabinowitz and Eikeland (1964) to group test items into halves or strata to create additive main and interaction effects, Huck proposed using a technique "invented" by Tukey to partition sums of squares (SS) for the residual to identify the nonadditive component where
\[SS_{\text{nonad}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{K} X_{ij}(\bar{X}_{ij} - \bar{X}_{..})(\bar{X}_{ij} - \bar{X}_{..})^2}{\sum_{i=1}^{N} (\bar{X}_{ij} - \bar{X}_{..})^2 \sum_{j=1}^{K} (\bar{X}_{ij} - \bar{X}_{..})^2}.\]

Then \(SS_{\text{nonad}}\) was subtracted from \(SS_{\text{res}}\) to leave \(SS_{\text{balance}}\). Since there was one degree of freedom associated with \(SS_{\text{nonad}}\), the formula for \(MS_{\text{bal}}\) became

\[MS_{\text{bal}} = \frac{SS_{\text{bal}}}{(N-1)(K-1)-1}.\]

Huck's estimating formula for reliability became

\[r = \frac{MS_{\text{exam}} - m' MS_{\text{bal}}}{MS_{\text{exam}}}.\]

Only one example was presented by Huck (1978) using data that was not dichotomously scored. In the example with three subjects and four items, reliability was estimated to be .629 by Winer's modification of Hoyt's ANOVA and .706 with Huck's method. No further comparison was made.

Alternatives to the Kuder-Richardson Model

Alternatives to the Kuder-Richardson model have been proposed. Cliff (1984) suggested a \(\phi\) alternative. Refinements and reconceptualizations of the basic model were
suggested (such as Guttman, 1945; Huck, 1978; and Winer, 1971). Other individuals (such as Ayabe, 1994; Callender and Osburn, 1979) developed unique solutions.

The Phi Coefficient

Kuder and Richardson's formula 20 was advocated by the developers as a useful formula for the estimation of reliability in circumstances where items difficulties were allowed to vary, but inter-item correlations were assumed to be equal. Johnson (1945) demonstrated that these two propositions were contradictory. The only circumstance where inter-item correlations could be equal was when all items were of equal difficulty. According to Loevinger (1947), in such a circumstance, passing one item implied that all items would be passed since they were of equal difficulty. Failing one item would result in failing all items. Loevinger concluded that a test could have perfect reliability using KR20 only in this "absurd instance" (Loevinger, 1947, p. 11).

The KR20 model is based on the phi ($\phi$) coefficient as an index of inter-item correlation. It is a variation of the fundamental equation for correlation, $r = \Sigma xy / N\sigma_x\sigma_y$, and related to Chi-square ($X^2$) such that $X^2 = N\phi^2$. Phi is intended for use as a measure of association between two dichotomous variables. By definition, the range of $\phi$ is from +1 to -1; however, $\phi$ can only achieve its maximum value when the proportions in corresponding categories are equal.
When category proportions are unequal, the range of $\phi$ is restricted.

Johnson (1945) advocated the use of the ratio of the actual $\phi$ to the maximum possible $\phi (\phi/\phi_{max})$ as an indicator of how much a test could be improved. If the inequalities between item difficulties were too great, although the test was measured as far from perfect, it may be as good as could be devised under the circumstances.

Brogden (1946) explored ramifications of the restricted range of $\phi$. He compared reliability estimates as computed by KR20 and KR2. He used KR2 as a standard because it has been accepted as the last of the Kuder and Richardson equations not requiring assumptions about item difficulty or factorial homogeneity (Brogden, 1946). He viewed these two assumptions as inter-related since "it is apparent that the assumption that they [the intercorrelations of items] can be accounted for by a single factor cannot be met unless the items are of equal difficulty" (Brogden, 1946, pp. 517-518). According to Brogden (1946), the assumption of factorial homogeneity was used only as a means of estimating the diagonal entries of the matrix in the numerator and it is not immediately evident how significant a failure resulted from the inability to meet the assumption.

To evaluate the effect of violating the assumption of equally difficult items, Brogden created a series of tests varying in length from 9 to 153 items with assumed
tetrachoric correlations from .2 to .8. The proportion of correct items ranged from .03 to .97, with three types of distributions (skewed, rectilinear, and normal). Differences in reliability estimates between the two methods increased when (a) the item intercorrelations increased, (b) the number of items decreased, or (c) there was a rectilinear distribution of items. The largest difference between KR2 and KR20, .088, was observed in the 9-item test with a rectilinear distribution of scores and a tetrachoric correlation of .8. Brogden (1946) concluded that KR20 was not influenced seriously by the difficulty bias in the product-moment intercorrelations. In the exceptional cases where bias was apparent, the spread in item difficulty and the degree of assumed item intercorrelation was greater than what typically occurs in practice. Published tests for practical use rarely have tetrachoric correlations above .30 and the typical range for item difficulties is .50 to .99 (Cronbach, 1951).

Cronbach (1951) explored the change in $\phi$ associated with changes in item difficulty by holding constant the relationship between 'underlying traits' and fixing the tetrachoric correlation. When the tetrachoric correlation was .30 the difficulty of one item was .50 and the other allowed to vary from .10 to .90, $\phi_{ij}$ ranged from .14 to .19. As the tetrachoric correlation increased, so did the range of $\phi_{ij}$. Cronbach concluded that the alleged limitation of $\phi$
had no practical affect for items of the sort used in psychological tests.

To evaluate Johnson's (1945) use of the ratio of the actual $\phi$ to its maximum possible value, Cronbach computed $\frac{\phi_{ij}}{\phi_{ij(\text{max})}}$ for the same data, and found that it was affected by variation in item difficulty more than $\phi$. While the range of $\phi$ was from .14 ($p_i = .10$) to .19 ($p_i = .50$), $\frac{\phi_{ij}}{\phi_{ij(\text{max})}}$ ranged from .19 ($p_i = .50$) to .42 ($p_i = .10$). Cronbach concluded that the indices intended to be substituted for the average $\phi$ were affected more severely than $\phi$ by differences in difficulty.

Cronbach prepared four hypothetical 45-item tests with normal, peaked, and rectangular score distributions to further substantiate his position that variations in $\phi$ are small within the range of psychological tests. Tetrachoric correlations were fixed at .30 in each case. Item difficulties ranged from $p_i = .10$ to $p_i = .90$. Comparing tests, the greatest difference in $\phi$ was .039 which resulted in a change in $\alpha$ from .892 to .914 (.022). He concluded that the effect on $\alpha$ was not important.

Finally, Cronbach (1951) examined a "perfect scale" where all $p_i$'s were equal to $\phi_{\text{max}}$. Items on the five item test had difficulty levels of .50, .58, .71, .80, and .89. Correlations ranged from 1.00 to .36. In this test of five items, $\alpha = .86$. Cronbach then increased the test size by increasing the number of items at each level. When the
number of items was 10, $\alpha = .951$ and for 20 items, $\alpha = .977$. According to Cronbach (1951):

It follows that even if items are much more homogeneous in content than the present tests and much freer from error, the cumulative properties of covariance terms makes the failure of all $\phi$'s to reach unity of next-to-no importance. $\alpha_{max}$ would be lower if difficulties range over the full scale, but the same principle holds. $\alpha$ is a good measure of common-factor concentration, for tests of reasonable length, in spite of the fact that it falls short of 1.0 if items vary in difficulty.

(p. 327)

A Phi Alternative

Cliff (1984) agreed with the fundamental reasoning of Kuder and Richardson (1937); however, he believed Kuder and Richardson's use of the average inter-item covariance as an estimate of the covariance between an item in the actual test and the hypothetical second test resulted in an underestimation of parallel form reliability. He argued it was reasonable to believe that two supposedly parallel items probably had a greater correlation than two items not intended to be parallel: potentially differing in content, factor structure, or difficulty.

Cliff (1984) proposed the use of the average Goodman-Kruskal gamma ($\gamma$) as an alternative to $\phi$. According to
Cliff (1984), γ is not influenced by item difficulty. The average γ should therefore be a more stable estimate than ϕ.

Gamma is a form of Kendall's tau, a monotonic function of Guilford's "cosine-pi" approximation to tetrachoric correlation, and not sensitive to ties. It is calculated from two-by-two tables of item pairs such that

$$\gamma_{jk} = \frac{(ad - bc)}{(ad + bc)}$$

where a, b, c, and d are the proportions of individuals returning the answer patterns correct-correct, correct-incorrect, incorrect-correct, and incorrect-incorrect for the items j and k.

Cliff (1984) recommended the calculation of the average γ rather than direct calculation of individual γ's to avoid the influence of extreme item splits. The average γ is defined as

$$\bar{\gamma} = \frac{\sum (ad - bc) / n}{\sum (ad + bc) / n}.$$  

Knowing γ and the item difficulties, the covariance of the item and its counterpart can be estimated as
According to Cliff, if $\gamma = 1$, then $s_{jj'} = (p - p^2)$, and consequently, in a perfect Guttman scale, the predicted $r_{xx'} = 1.0$. The estimated item covariances can then be substituted into the conceptual form for reliability such that

$$
\hat{r}_{xx'} = \frac{s_x^2 - \Sigma s_j^2 + \Sigma s_{jj'}}{s_x^2}.
$$

Cliff presented the use of $r_\gamma$ two ways. To illustrate the method of computation, a 3-item test was used. Obtained reliability estimates were $r_\gamma = .581$, KR20 = .517, Horst's reliability = .754, and $\phi/\phi_{\text{max}} = .625$.

Cliff also simulated the use of $r_\gamma$ and KR20. He sampled items from a population with known item discrimination and difficulties and computed $r_\gamma$, KR20, and parallel form reliability ($r_{xx'}$) under conditions of varying difficulty, consistency, and test length. With tests of 5 and 10 items and samples of 100 subjects, the typical order of estimates was $r_{xx'} > r_\gamma > \text{KR20}$, although the differences among estimates were not remarkable. The average difference
between $r_{xx}$ and $r_\gamma$ was .007. The average difference between KR20 and $r_{xx}$ was .047.

Cliff repeated part of his study with a sample size of 500 subjects and additionally calculated the Pearson correlation coefficient between the total score and the true scores. The average difference between $r_{xx}$ and $r_\gamma$ was .024, with $r_\gamma$ being higher. The difference between estimates was reported to be not significantly different (α not specified). The KR20 value was always lower than $r_{xx}$, by an average of .051. The difference was reported to be significantly different from zero (level of significance not specified). Cliff (1984) observed that KR20 was less of an underestimate of reliability when there were more items. He concluded that $r_\gamma$ could be used to estimate $r_{xx}$ when a realistic upper bound to the parallel form reliability was desired. The difference between KR20 and $r_\gamma$ was expected to be quite small except under extreme conditions such as with less than 10 items or if there was appreciable variability in difficulty ($s_p$ of .15 or more) (Cliff, 1984).

According to Cliff (1984), the greatest difficulty of $r_\gamma$ seemed to be the lack of a formal estimation procedure. Derivation of estimation procedures would be useful.

Cliff concluded that $r_\gamma$ was an alternative to KR20 without the unrealistic assumption that average covariance of equally difficult items be the same as the average covariance of items of unequal difficulty. He argued $r_\gamma$
should be viewed as an upper limit to parallel form reliability since it assumed exactly parallel distribution of item difficulties. No assumption of homogeneity was made in the derivation (Cliff, 1984).

Unlike the work of Cronbach (1951), Horst (1953), Kuder and Richardson (1937), and Loevinger (1947), there is no evidence of comment on Cliff’s proposal in the literature. If one accepts Kuder and Richardson’s conceptualization of reliability estimation as correct, and that \( \gamma \) is unaffected by the limitations of the \( \phi \) correlation; then Cliff’s \( r_\gamma \) is a substantial contribution to psychometrics.

Three comments can be made concerning Cliff’s investigation. First, the use of small numbers of test items in Cliff’s examples suggest the cumbersome nature of computations; although a computer routine for the calculation of \( r_\gamma \) should be possible.

Secondly, Cliff seemed to use parallel forms as the standard for comparison. It is more reasonable that parallel form reliability is a step toward gaining information about the population. It is unfortunate that Cliff did not focus more on comparing \( r_\gamma \) to \( r_m \) instead of \( r_m \), since his experimental design presented one of the rare circumstances when a good estimate of the population parameter was known.

Finally, Cliff’s data revealed that \( r_\gamma \) frequently yielded higher reliability estimates than did \( r_m \). The
investigator did not state whether \( r \) ever exceeded \( r_m \) in his large sample study. The usefulness of any lower bound, even an 'upper' lower bound would be diminished if there were circumstances when the estimate exceeded the true reliability.

**Guttman's Lower Bounds**

Guttman (1945) derived a series of lower bounds for reliability, including the Kuder-Richardson formula. The formulations presented by Guttman differed from convention in five ways: (a) variation over trials, persons, and items were kept distinct, with unreliability defined as variation over trials; (b) the reliability coefficient was defined without assuming zero correlation for error, provided that the total variance of the test remained the sum of error and expected scores; (c) it was emphasized that a coefficient of reliability could not be calculated from a single trial, instead, there was a focus on what information could be obtained without repeat measurement, most specifically, the determination of lower bounds for the estimate of reliability; (d) one basic assumption was that errors of observation are independent between items and persons over the universe of trials; and (e) no assumptions were made about relationships between the items themselves (if without experimental error).

Six lower bounds \((\lambda_1-\lambda_6)\) and their computing formulas \((L_1-L_6)\) were presented by Guttman to be used with large
samples and only cautiously with smaller samples. The most simple lower bound, $\lambda_1$, represented the ratio of off-diagonal covariances to the total test variance. The computing formula for $L_1$ was:

$$L_1 = 1 - \frac{\sum s_j^2}{s_x^2}.$$ 

The second lower bound, $\lambda_2$, increased the value of the $L_1$ estimate by adding the sum of squares of the covariances between items $(C_2)$, such that

$$L_2 = L_1 + \frac{\sqrt{\frac{K}{(K-1)}} C_2}{s_x^2}.$$ 

To save the labor of computing covariances, Guttman proposed $\lambda_3$, which was computed as

$$L_3 = \frac{K}{K-1} L_1.$$ 

The $L_3$ equation was computationally equivalent to KR20; however, the meaning attributed to the value differed between Guttman and Kuder and Richardson. Guttman proposed
λ₃ as a lower bound to reliability while KR20 represented a computing formula for the estimation of reliability. According to Guttman (1945), the intended use of λ₃ did not require meeting the stringent assumptions required of KR20. According to Guttman, L₃ was so easy to compute that L₄ need not be used as an estimate. These first three lower bounds had a hierarchal relationship where $L_1 < L_3 < L_2$.

Guttman's $\lambda_4$ was a corrected split-half correlation coefficient such that

$$L_4 = 2(1 - \frac{S_a^2 + S_b^2}{S_x^2}),$$

where $S_a^2$ and $S_b^2$ represent the test variances for two partial splits of a test. Regardless of which split was used, $L_4$ remained a lower bound to reliability (Guttman, 1945). Jackson and Agunwamba (1977) believed that $\lambda_4$ provided the highest estimate of reliability at that time.

Guttman's fifth and sixth coefficients were for use in special situations. When one test item had larger covariances with other items as compared to the covariances among other items, $L_5$ would produce a higher estimate of reliability; otherwise the $L_5$ estimate would be lower. If test items had high multiple correlations but low zero-order intercorrelation, $L_6$, which was based on multiple
correlation, would produce an estimate at least as high as $L_2$. In other circumstances, the $L_6$ coefficient would be less than $L_2$.

An exploration of the relationship between KR20 and Guttman's $L_2$ was conducted by ten Berge and Zegers (1978). They observed that $L_2$ was at least as good an estimate of test reliability as KR20. The superiority of $L_2$ was attributed to the use of the sum of squares of covariances between items as well as the sum of the covariances. Ten Berge and Zegers (1978) demonstrated that the sum of squares could be raised to higher powers without exceeding true reliability. The investigators compared a series of $\mu$ coefficients, where $\mu_0$ was equivalent to KR20 (Guttman's $L_2$), $\mu_1$ was Guttman's $L_2$, $\mu_2$ was Guttman's $L_2$ with the sum of squares squared again, such that

$$
\mu_2 = \frac{1}{\sigma^2} \left( \sum_{i \neq j} \sigma_{ij}^2 + \left( \sum \sigma_{ij}^2 \left( \frac{n}{n-1} \sum \sigma_{ij}^4 \right)^{1/2} \right)^{1/2} \right).
$$

To illustrate the use of $\mu_2$, ten Berge and Zegers reported data from eight college tests on various psychology topics. No information was provided concerning number of items or number of subjects. Differences between $\mu_0$ and $\mu_1$ ranged from .020 to .076. As powers of $\mu$ increased, the difference in estimates decreased. Differences between $\mu_1$
and $\mu_2$ ranged between .001 and .006. Between $\mu_3$ and $\mu_6$ there was an increase in the reliability estimate for only one test. The difference was .0003. The investigators concluded that computing $\mu_1$ was beneficial, but further computation was not worthwhile.

With very few items, there was more of an increase in the reliability coefficient as higher powers of covariance were added. To illustrate, the authors recomputed reliability estimates using only the first three and four items of each test. They presented estimates from a test described as "showing the best results for $\mu_2$" (ten Berge & Zegers, 1978, p. 579). For the 3-item test, increases in reliability coefficients were .050, .019, and .005 between $\mu_0$, $\mu_1$, $\mu_2$, and $\mu_3$ respectively. Differences in the 4-item test were smaller: .024, .007, and .002. Ten Berge and Zegers (1978) concluded that for tests with a wide spread in $\sigma_j$, the calculation of $\mu_2$ or $\mu_3$ may be beneficial. But such tests rarely occur in practice, therefore it is generally sufficient to calculate $\mu_1$.

**The Greatest Lower Bound**

An attempt to quantify the uppermost value of the lower limit of reliability or "greatest lower bound" (GLB) was conducted by Woodhouse and Jackson (1977). In situations where the observed score dispersion matrix was assumed to be known or well-estimated, they used a computer search procedure to find the greatest lower bound. Three
artificial data sets with a sample size of three subjects were presented. In one test there were no differences in estimates computed by KR20, $\lambda_2$, and GLB, but the $\lambda_4$ estimate was .102 lower. In the other two tests, the value of $\lambda_4$ exceeded KR20 and $\lambda_2$, but not GLB. They concluded that $\lambda_4$ was not always the greatest lower bound, but it was a better estimate than $\lambda_2$ or KR20. The authors emphasized that their computer search procedure was valid only when the observed score dispersion matrix was assumed to be known or well-estimated, such as in circumstances where a national test was administered by major testing corporations. Before their technique could be used otherwise, more information about the sampling distribution of test scores was required.

An Algorithm for Estimating the Maximum Split-Half

In 1977, Callender and Osburn presented a method for estimating reliability by approximating the largest possible split-half reliability coefficient. Their work was based on Guttman's (1945) demonstration that the lower bound for true reliability of a composite was the largest of any of the possible split-half coefficients. Recognizing that computation of the $K! / \{2[K/2]!\}$ possible splits computationally unmanageable as the number of items increases, they sought an algorithm for discovering a large split-half without computing all the alternatives. Callender and Osburn presented the split-half reliability formula
where $\sigma_{AB}$ refers to the covariance between test halves A and B. Regardless of split, test variance ($\sigma^2_x$) remains constant, therefore the magnitude of reliability is influenced directly by the covariance between two test halves. The covariance of the two test halves equals the sum of every item on test A with test B, thus, the goal of the MSPLIT algorithm was to assign items to test halves in such a way as to maximize the covariances.

An algorithm was devised to produce a new covariance matrix, half the size of the original, thereby reducing the number of computations by limiting the size of the matrix. The first step was to examine the original item covariance matrix and identify the two items with the largest covariance. The two items with the largest covariances served as a starting point for the new compact matrix. In the second step, the remaining items were evaluated in combination with the already placed items to determine the next set of items to be included in the matrix. The process continued until all the items were arranged in the matrix (Callender & Osburn, 1977).

To illustrate the difference between the MSPLIT and the actual largest split-half correlation coefficient, Callender
and Osburn (1977) compared results in 16 samples from each of nine 10-item tests across a range of reliabilities. The set with the smallest split-half reliability coefficient of .315 had an MSPLIT of .295, a difference of .020. As test reliability increased, MSPLIT increased and there was a general trend toward a decreased difference between split-half and MSPLIT coefficients. The most reliable item set had a split-half coefficient of .785 and MSPLIT coefficient of .776, a difference of .009. Overall 144 samples were evaluated. In 41% percent of the samples, the difference between the two methods was less than .001. In another 29%, the difference was less than .021. Ninety-eight percent of the samples had differences of less than .081 (Callender & Osburn, 1977).

Callender and Osburn (1977) also calculated odd-even split-half and KR20 reliability estimates on the same data. The mean MSPLIT's exceeded the odd-even and KR20 estimates in all nine item sets. The differences in these estimates for each of the data sets were not reported. Comparing the means, the estimate obtained by MSPLIT was always the largest and KR20 was lowest except for one data set. The mean of differences between estimate means was .106 between MSPLIT and odd-even split halves and .116 between MSPLIT and KR20.

Recognizing that the statistical maximizing process may provide inflated estimates because of sampling variation,
the same investigators compared reliability estimates estimated from samples of 100 subjects on nine 10-item and three 40-item tests (Callender & Osburn, 1979). Callender and Osburn concluded that the MSPLIT estimated the largest split-half coefficient more accurately than KR20. Although the MSPLIT consistently overestimated the largest split in the 40-item tests, "this does not necessarily imply that the population reliability was overestimated, since the MSPLIT population coefficient may not have been the largest $\lambda_4$" (Callender & Osburn, 1979, p. 96). The investigators also remarked about the lower estimate of reliability by KR20, noting that the extent of difference in estimates decreased as the estimated test reliability increased. They concluded that the MSPLIT was a better estimator of reliability, noting that overcorrection for attenuation would be the result of using other estimates. In addition, since internal consistency indicators may be used to infer an experiment or statistic's power, an underestimate of reliability would lead to an underestimate of power (Callender & Osburn, 1977).

It should be questioned whether the intention of an estimate is to provide as close a guess of the true reliability irrespective of over- or underestimation or a directional guess of true reliability. Given an MSPLIT estimate, it is unclear whether that value represents an over- or underestimate of the true reliability. By making a
directional guess, one is able to make some inference regarding the lower bound. As long as the reliability estimation procedure yields a value at or below the true reliability, then it is possible to make decisions based on the obtained information. If one accepts the notion that the best estimate of true reliability in the Callender and Osburn study was the population $\lambda_4$, then comparing the obtained values for the MSPLIT from the samples to those standards, in the 10-item tests, eight of the nine MSPLIT's exceeded their respective population $\lambda_4$'s (by as much as .117). None of the means of KR20 estimates exceeded the population $\lambda_4$. Admittedly, the mean KR20 was an underestimate of the $\lambda_4$ by as much as .296; however, the interpreter of the KR20 data should be reasonably confident that true reliability is somewhere above the estimated KR20 value. The interpreter of MSPLIT may anticipate proximity to the true reliability, but has no way to predict whether the calculated value is above or below the true value.

Homogeneity: A Related Concept

An index of homogeneity for tests of ability was proposed by Loevinger (1947). Ability was defined as "the immediate possession of achievement" (Loevinger, 1947, p. 27). She made two assumptions. First, that it was possible for an examination to test the same ability at different levels. Secondly, the possession of abilities required to answer one test item would not impede performance on any
other item. Loevinger defined a test as perfectly homogenous if "A's score is greater than B’s score, then A has more of some ability than B and it is the same ability for all individuals A and B who may be selected" (Loevinger, 1947, p. 28). In a perfectly heterogeneous test each item measured an ability independent of the abilities measured by the other items.

Two theorems were proposed. Loevinger's (1947) first theorem stated "when the items of a perfectly homogenous test are arranged in order of increasing difficulty, every individual will pass all items up to a certain point and fail all subsequent items" (Loevinger, 1947, p. 28). The first theorem allowed her to operationalize a definition of homogeneity by making a series of assumptions. If a test was perfectly homogenous and items were arranged in order of increasing difficulty, then if an item was known to be passed, all previous (easier) items were also passed. In a perfectly heterogeneous test, the probability of passing any given item was independent of the knowledge about passing any other item. Secondly, "the probability of passing any item A for those known to have passed any other item B is not less than the probability of passing A for those whose response to item B is not known" (Loevinger, 1947, p. 29). Expressed in terms of probability, \( P_{ij} = P_{ij} / P_j \) where \( P_j \) is the probability of passing item \( j \), \( P_{ij} \) is the probability of passing items both \( i \) and \( j \), and \( P_{ij} \) is the probability
of passing item $i$ given $j$. Applying the second assumption, $P_{ij} \geq P_i$. By definition, in a perfectly heterogeneous test, $P_{ij} = P_i$. Also by definition, item $j$ is considered more difficult than item $i$ if $P_j < P_i$. When test items are arranged in order of increasing difficulty, a perfectly homogeneous test may be defined where $P_{ij} = 1$ for all $j > i$.

Loevinger wrote:

(Apparently the degree of homogeneity of a test depends on the values of the quantity $P_{ij}$ for all pairs of items, and an adequate index of homogeneity must be based on all of these values. For each pair of items, $j$ greater than $i$, the quantity $P_{ij}$ has a value between a lower limit of $P_i$ and an upper limit of unity. It is natural therefore to build an index of homogeneity around the quantities $P_{ij} - P_i$. (Loevinger, 1947, p. 30)

While the probability of success on a given item $p_i$ was a characteristic of the population, the probabilities of $P_{ij}$ were a characteristic of the sample, depending on the magnitude of $p_j$. Loevinger proposed weighting her index of homogeneity according to $p_j$. Loevinger's index, $H_i$, was the ratio $S/S_{\text{max}}$, where

$$S = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} P_i (p_{ij} - p_i)$$

and
In a perfectly heterogeneous test, \( P_{ij} = p_i \) and \( S = 0 \) (Loevinger, 1947). Whenever \( p_{ij} \geq p_i \), \( S \) must always be either 0 or positive since each individual term is 0 or positive. Loevinger noted that the maximum value of \( S \), or \( S_{\text{max}} \), becomes unity when \( p_{ij} \) is maximized. She believed that the characteristics of \( H_t \), with a potential range of values from 0 to 1, were ideal for an index of homogeneity.

Loevinger's second theorem stated that \( H_t \), the coefficient of homogeneity, was a linear function of the test's variance, with the difficulties of the items defining the functional constants. Because of this theorem it was possible to demonstrate that \( H_t \) could be defined in terms of the population variance and the probabilities of passing successive items. Loevinger's index could be described as

\[
H_t = \frac{V_x - V_{\text{het}}}{V_{\text{hom}} - V_{\text{het}}}
\]

where \( V_x \) was the population variance, \( V_{\text{het}} \) was the variance of a perfectly heterogeneous test with the same distribution of item difficulties, and \( V_{\text{hom}} \) was the variance of a perfectly
homogeneous test with the same distribution of item difficulties.

Loevinger estimated $H_i$ from sample statistics by substituting the sample variance for the population variance, the proportion of individuals passing an item in the sample for the comparable proportion in the population, and assuming that the order of difficulty for the population was the same as that observed in the sample. Following these substitutions and algebraic simplification, the index of homogeneity could be estimated as

$$Est \, H_i = \frac{N(\Sigma X_k^2 - \Sigma X_k) + \Sigma N_i^2 - (\Sigma X_k)^2}{2N(\Sigma iN_i - \Sigma X_k) + \Sigma N_i^2 - (\Sigma X_k)^2}$$

where $N_i =$ number of individuals passing the $i$th item

$\Sigma X_k =$ sum of raw scores for all $N$ individuals.

Loevinger cautioned that the sampling properties for the estimate of homogeneity had yet to be determined and that the magnitude of expected sampling error under various conditions was unknown. She recommended that the use of $H_i$ should be restricted to samples of at least 100 cases until these characteristics were known. Loevinger (1947) provided no examples of using her estimate or its relationship to KR20.
Horst (1953) agreed with Loevinger that KR20 could yield unity as an estimate of reliability only when all items were of equal difficulty. This was true because $\phi$ could equal unity only when items were of equal difficulty, and Horst cited Loevinger's argument that unless a homogeneous test contained items of differing difficulties, a single test item would be as discriminating as the entire test.

Horst disagreed with Loevinger that $\phi$ should be used as an index of internal consistency. He thought Loevinger stopped short of her intended goal of developing an index of internal consistency. He believed that $\phi$ was an estimate of item reliability and needed to be inserted into Kuder and Richardson's third equation where

$$ I_{tt} = \frac{S_X^2 - \sum_{i=1}^{K} pq + \sum_{i=1}^{K} I_{ii} pq}{S_X^2} . $$

This early Kuder and Richardson formula represented a conceptualization prior to the making of assumptions other than a unidimensional test. Horst rearranged the terms of KR3 so that
Horst then used a reorganized version of Loevinger's $H_i$ as an estimate of item reliability ($r_{ii}$) resulting in the formula

$$r_{tt} = \frac{S_x^2 - (1 - r_{ii}) \sum pq}{S_x^2}$$

which simplified to

$$r_H = \frac{S_x^2 - \Sigma pq}{S_m^2 - \Sigma pq} \frac{S_m^2}{S_x^2},$$

where $S_m^2$ was the maximal test variance and

$$S_m^2 = 2 \Sigma ip_i - M_x(1 + M_x),$$

where $\Sigma ip_i$ reflected the sum of products of item difficulties multiplied by their respective rank when arranged in order of descending magnitude and $M_x$ was the mean score of the test.
According to Horst (1953), his estimate provided a more realistic estimate of the reliability than KR20. Even in circumstances where item difficulties varied, the upper limit of reliability was unity. With the addition of $S_m^2/S_x^2$, the value reported by this estimate will always be higher than Loevinger's index.

The principal disadvantage of Horst's method was the computation of $S_m^2$, which required the calculation of $\Sigma ip_i$. Compared to KR20, calculation of $\Sigma ip_i$ was an additional step, although the more significant portion of the computation was accomplished in the process of calculation item difficulties, a requirement for both KR20 and the Horst method (Horst, 1953).

Horst compared his method to KR20 using one example. He presented a 7-item test with item difficulties in a rectangular distribution ranging from .20 to .80. The mean test score was 3.5 (50% correct overall). By KR20, reliability was estimated as .84. Using his method, reliability was .92. Horst observed that the difference in calculated reliabilities varied with the range of item difficulties. Tests with a broader range of item difficulties had a greater disparity between the two methods.

Horst believed his method, like KR20, was to be considered a lower bound for reliability. Compared to KR20, his lower bound would be "somewhat higher" (Horst, 1953, p.
Horst emphasized that his method corrected for attenuation from the dispersion of item difficulties, but it still required the test be unidimensional and assumed item unreliability was the only source contributing to failure to achieve maximum item intercorrelations.

**Criticisms of Loevinger and Horst**

Loevinger's objection to the Kuder-Richardson model was the inability of $\phi$ to achieve its maximum value of unity unless the two items being compared had equal difficulties. To devise a scale with a maximum value of 1.00, Loevinger divided $\phi$ by $\phi_{ij}(\text{max})$. Horst reasoned that $\alpha$, like $\phi$, must be less than 1.00 when item difficulties are not equal. He rescaled $\alpha$ in a fashion similar to Loevinger's treatment of $\phi$. There was no mathematical justification for the use of ratios (Cronbach & Azuma, 1962; Guilford & Fruchter, 1973).

Horst's method was tested by Cronbach and Azuma (1962). They hypothesized Horst's method, like Loevinger's index, would be sensitive to item difficulty. Cronbach and Azuma created hypothetical parallel tests by sampling randomly from a defined universe of items. The tests had uniform tetrachoric intercorrelations and were assumed to be unidimensional.

Seven estimates of internal consistency were evaluated including parallel forms ($\rho_{TT}$); correlation of the random test with true score ($\rho_{TM2}$), which was defined as the proportion of observed score variance that could be
accounted for by the score by the person across the universe of items (this estimate is not available in actual practice since universe parameters are unknown); \( \alpha \) (KR20), \( \alpha_s \) as introduced by Jackson and Ferguson (1941) for use with tests that could be stratified on the basis of content or difficulty; \( \alpha_L \) for item-parallel tests where items were matched on the basis of difficulty and content (Lord, 1955); \( \alpha_H \), or Horst's corrected \( \alpha \); and KR21. Estimates were computed for 9- and 54-item tests with tetrachoric correlations \((r_{ij})\) of 1.00, .70, and .30. Higher values of \( r_{ij} \) (.70 and 1.0) were of theoretical interest since the behavior of \( \phi \) is more anomalous as \( r_{ij} \) increases. Higher intercorrelations are most often encountered in Guttman scales for attitude measurement in tests used for research purposes. Results when \( r_{ij} \) was .30 and below were of more importance. It is rare for a published test with practical use to have tetrachoric correlations in the range above .30 (Cronbach & Azuma, 1962).

Item difficulties were uniform from .01 to .99. The investigators observed that such a wide distribution was more likely to be found in tests of personality than ability. They also limited the range of item difficulties to .50 to .99, more typical for ability tests. Cronbach and Azuma (1962) predicted that the wide range of item difficulties in combination with small sample sizes should influence peculiarities of \( \phi \).
For the nine-item test, with $r_{ij} = .3$, the mean estimated $\rho_{TM2}$ and $\rho_{TT}$ were .613 (.55-.67) and .609 (.58-.65), $\alpha$ was .599 (.51-.67), and Horst's index was .668 (.63-70). With $r_{ij} = .7$, the mean estimated $\rho_{TM2}$ and $\rho_{TT}$ were .839 (.74-.87) and .836 (.80-.87), $\alpha$ was .819 (.74-.88), and Horst's index was .915 (.90-92). With $r_{ij} = 1.0$, the mean estimated $\rho_{TM2}$ and $\rho_{TT}$ were .920 (.75-.98) and .908 (.79-.98), $\alpha$ was .895 (.82-.96), and Horst's index was 1.00. Cronbach and Azuma (1962) concluded that $\alpha$ underestimated the expected $\rho_{TM2}$ and $\rho_{TT}$ slightly, with the average difference being .01 or .02. Horst's index yielded coefficients that were higher than other formulas by as much as .05. Differences between estimates from the 54-item test were smaller, but consistent. Again $\alpha$ slightly underestimated the expected $\rho_{TM2}$ and $\rho_{TT}$ while Horst's index exceed them, but by less than .01.

When discussing Horst's estimate, the investigators observed that the range of estimates exceeded the expected value, particularly on the 9-item test. According to Cronbach and Azuma (1962):

Horst proposed $\alpha_H$ as a "realistic estimate of the reliability" defined as "consistency of behavior within a very limited time interval." Neither Horst nor Loevinger rationalized the proposal to divide the obtained coefficient $\alpha$ or $\phi$ by its maximum value, nor did Horst give any empirical justification for
considering \( \alpha_H \) to be "realistic." Since his aim was to remove effects associated with the dispersion of item difficulties, and since \( \alpha_L \) ([item parallel]) accomplishes this, we see no merit in his proposed coefficient which greatly exceeds \( \alpha_L \). (p. 659)

Cronbach and Azuma (1962) recommended computation of \( \alpha \) unless the tetrachoric correlation was believed to be unusually high, when \( \alpha_s \) should be used. They also observed that when the number of test items was small, substantial variable error occurred. The investigators restricted their findings to single common factor tests:

When a test is a random sample from a pool of dichotomous items whose content represents a single factor, and whose intercorrelations are within the normal range, the \( \alpha \) coefficient is highly satisfactory as an estimate of the inter-test correlation and the percentage of variance attributable to true score.

(Cronbach & Azuma, 1962, p. 663)

The Ratio of Actual to Maximal KR20 Estimates

Maximum and minimum values for KR20 were explored by Terwilliger and Lele (1979). They agreed with Tucker (1949), that KR20 was at its minimum when all items were of equal difficulty, and this minimum value was KR21. The relationship between KR20 and KR21 was illustrated as
\[ KR20 = KR21 + \frac{K^2}{K-1} \left[ \frac{S_p^2}{S_x^2} \right] \]

where \( S_p^2 \) represents the variance of item difficulties. Kuder and Richardson formula 20 will be at a maximum when \( S_p^2 \) is greatest, assuming a constant total score variance. Terwilliger and Lele (1979) provided a mathematical proof that the maximum value for the variance of item difficulties occurred only when items formed a Guttman scale, assuming a fixed score distribution. The disparity between the maximum and minimum values was greatest when the test difficulty was .5. As the test became either easier or more difficult, the difference between maximum and minimum values decreased (Terwilliger and Lele, 1979).

Terwilliger and Lele (1979) also explored the effect of the distribution of test scores on the minimum and maximum values for KR20. Distributions with greater variance in test scores had higher maximum values of KR20. The investigators concluded that both difficulty level and distribution of test scores are important considerations related to internal consistency. Tests with extreme levels of item difficulty and large variances virtually assure positive internal consistency, especially if the distribution of test scores is rectangular or bimodal (Terwilliger & Lele, 1979).
Raju (1982) expanded on the work of Terwilliger and Lele (1979). He observed that "Loevinger and Horst defined the maximum test variance under the condition that the item difficulties remain the same, or alternatively that the item covariance be maximum when the item difficulties are fixed" (p. 150). Raju, like Terwilliger and Lele, emphasized maximizing item covariances when test variance was fixed, but the item difficulties were allowed to vary. The following formula was offered:

\[
I_R = \frac{S_X^2 - \sum_{i=1}^{n} Q_i}{S_X^2 - \sum_{s=1}^{n} Q'_s}.
\]

Raju’s formula represented the ratio of the sum of actual item covariances to the sum of maximum possible item covariances under the assumption that the total test variance remained the same.

As such, it is analogous to Loevinger’s index of homogeneity. Second, it is also the ratio of actual KR-20 to maximum possible KR-20 and therefore, it is also analogous to Horst’s index of homogeneity. Third, the proposed index attains the maximum value of one if [and only if] the items form a perfectly homogeneous test and the maximum is independent of the distribution of item difficulties. (Raju, 1982, p. 151)
To illustrate his formula's performance, Raju (1982) presented data from an eight item test administered to 10 individuals, with item difficulties ranging from 1.0 to .3 and scores ranging from 1 to 9. Errors were allowed. The range of estimates were: Guttman's Reproducibility .84, Raju's index ($r_R$) .64, Horst's index ($r_H$) .53, KR20 .49, and Loevinger's ($H_t$) .16.

Not all factors affecting $r_R$ are known. One important question was whether $r_R$ adequately measured test reliability. According to Raju, the answer is yes, using the arguments of Loevinger and Horst. Through the assumption that variance is fixed, $r_R$ is independent of the distribution of item difficulties, and can attain a value of 1 even when the item difficulties are not equal, thus $r_R$ becomes the ratio of actual to maximum KR20 (Raju, 1982).

Raju was unable to present a demonstration that his formula fit the classical definition of reliability (ratio of true score to total score variance), but:

it should be noted that the indices of Loevinger and Horst also face the same difficulty....

However, in the case where the test is perfectly homogeneous, it is fairly simple to demonstrate that formula (12) [$r_R$] and the classical definition of reliability are identical when the items are essentially tau-equivalent. (Raju, 1982, p. 152)
Semi-Order and Item Dominance Matrices

The concept of semi-order as a standard for estimating internal consistency was proposed by Cudeck (1980). Instead of relying on the score matrix, semi-order used item dominance matrices to compare the obtained dominance matrix with the expected matrix, building from a design such as the Guttman scalogram. It was anticipated that no individual would correctly answer a more difficult question once an easier item was missed and conversely, no one would incorrectly answer an easier item if a more difficult question were answered correctly. Departures from the expected matrix were reflected by two indices: $c_a$ which had a range from -1 to +1 and $c_g$ which ranged from 0 to +1. According to Cudeck (1980), $c_g$ was equivalent to Loevinger’s index of homogeneity.

To evaluate $c_a$, $c_g$, and KR20, Cudeck (1980) designed a Monte Carlo study using the three parameter logistic model of item response theory. The three parameter model incorporated the concepts of item discrimination, item difficulty, probability of chance success, with ability. Across various levels of these parameters, populations of 200 items and 300 subjects were generated, samples of 100 subjects on tests of 20 or 40 items were drawn, and internal consistency measures were computed. A total of 1280 values for each statistic was obtained. Cudeck (1980) believed the
range of parameters was representative, although not exhaustive.

Data were analyzed using 5-way ANOVA. The main effects were discrimination, ability, difficulty, chance, and sample size. Interaction effects were also analyzed. Numerical values were reported to be statistically significant at $p < .05$. No F-ratios or degrees of freedom were reported.

For $c_2$, the variance accounted for by main effects was .923. Main effect variance was partitioned as discrimination = .757, ability = .005, difficulty = .057, and chance .104. Two-way interaction accounted for 2.5% of the variance, of which 1.7% was attributed to the discrimination-by-chance interaction. No other source of variance contributed more than .5%. The total explained variance was .956.

For $c_6$, the variance accounted for by main effects was .891. Main effect variance was partitioned as discrimination = .662 and chance .228. Two-way interaction accounted for 8.0% of the variance, of which 7.4% was attributed to the discrimination-by-chance interaction. No other source of variance contributed more than .4%. The total explained variance was .977.

For KR20, the variance accounted for by main effects was .923. Main effect variance was partitioned as discrimination = .773, chance .066, and sample size .075. Two-way interaction accounted for 1.5% of the variance. No
source of variance contributed more than .6%. The total
explained variance was reported to be .914.

The reader's direction should first be directed to the
"reasonableness" of the KR20 numbers. Although the total
explained variance for \( c_2 \) is exactly the sum of main and
interaction effects, and similarly, the sum of effects for
\( c_3 \) differs by only .001 (presumably an rounding artifact).
The sum of main and interaction effects for KR20 is actually
.938, rather than .914 as noted. How the KR20 main effect
variance alone could exceed the total explained variance was
not explained (or noted) by the investigator.

The investigator described the study's significant
finding by focusing on the item discrimination parameter.
As mean discrimination increased, so did the reliability
index under study. For \( c_2 \) and \( c_3 \), the increase was linear
across the range of item discrimination from .25 to 2.0.
For KR20, the increase approached an asymptote. The value
of KR20 increased linearly with item discriminations from
.25 to .50. Above .50, the rate of increase diminished and
overall KR20 reliability increased only slightly once
discrimination reached values of approximately 1.00.

As the range of item difficulties was increased (from
\( \sigma_{diff} = 1.0 \) to \( \sigma_{diff} = 1.5 \)), \( c_2 \) appeared to increase by slightly
less than .20. With chance controlled at 0, there was the
predictable increase in consistency with increasing
discrimination. When a chance level of .2 was introduced,
the interaction with discrimination resulted in increasing attenuation of the improvement in $c_d$ as discrimination increased.

Cudeck (1980) reported that the main effects of discrimination and chance accounted for 85% of the variance in $c_d$ and $c_a$. It is unclear why he chose to highlight that fact when 84% of the variance in KR20 is attributed to the same two main effects.

Cudeck (1980) also found that neither KR20 nor $c_d$ was insensitive to variability in item difficulties. Cudeck's finding of insensitivity was in contrast to Cronbach (1951) who suggested that Loevinger's index may be affected more by item difficulty than KR20.

When comparing $c_a$ to $c_d$, Cudeck (1980) concluded that $c_a$ appeared not to be as useful as an index of internal consistency because it was less sensitive to chance and more sensitive to ability and difficulty. He concluded that $c_d$ appeared "quite satisfactory" (p. 124) on the basis of lack of effect by ability and difficulty and linearity across the range of discrimination. Cudeck's $c_d$ was also attractive because it was not dependent on the number of items, and required only modest theoretical assumptions.

Comparing $c_d$ to KR20, Cudeck (1980) presupposed that since both KR20 and $c_d$ attempt to assess the consistency of a score matrix, results from each measure should correspond. Doubling the number of items from 20 to 40 raised KR20 by 15
to 20%. Because change in KR20 was limited once item
discrimination exceeded .75, $c_9$ may be preferred in
situations with highly discriminating items. Cudeck (1980)
concluded that $c_9$ in the range of .3 to .5 corresponded to
"acceptable reliability in the usual senses" (p. 126) plus
provided additional information regarding consistency of
order in a Guttman scale sense. The investigator concluded
that $c_9$ was an acceptable estimate of consistency,
especially when the test designer was unwilling to make the
assumptions of classical test theory, when the number of
items is small. Cudeck's suggestion of using Loevinger's
index when the number of items is small contradicted
Loevinger's recommendation that the use of $H_t$ be limited to
tests with at least 100 items.

The linear responses of $c_a$ and $c_9$ to discrimination, as
demonstrated by the investigator, are definite strengths
compared to the curvilinear response of KR20. Beyond the
finding of linearity, Cudeck's data may be interpreted
differently. For the purpose of the alternative
explanation, it should be assumed that the values for
individual main and interaction effects were true as
presented by the investigator, and that the sums of the main
effects and total explained variance were errors in
typesetting. It would seem reasonable to strive for
robustness in the design of an index of consistency.
Ideally an index would reflect only one parameter:
discrimination. To the extent that other factors contribute to the estimate obtained, the utility of the index is compromised. Robustness can be evaluated by examining the amount of variance explained by item discrimination alone. Comparing variance accounted for by item discrimination, $c_a$ and KR20 are each superior to $c_a$. A second goal would be simplicity, that whatever affects the index, does so as a main effect, not interacting with other variables. The indices with least variance attributed to interaction are $c_a$ and KR20. With less interaction and slightly less alternative main effects, KR20 is the most useful index. Aside from linearity, the only positive aspect of $c_a$ is that it was affected by only one other main effect: chance. An increased likelihood of guessing the correct answer decreased the consistency estimate. Unfortunately, the effect of chance alone is obscured by the interaction between chance and discrimination. The graph presented by Cudeck (1980) illustrated the magnitude of the chance-discrimination interaction. Accepting Cudeck's definition of a reasonably consistent test, with a $c_a$ between .3 and .5, the minimum acceptable level of consistency (.3) was achieved with discriminations of approximately .5. With a chance level of .2, a .3 estimate of consistency was not achieved until item discriminations approached 1.0, a situation described by the investigator as "rare" (Cudeck,
1980, p. 126) except for special testing situations such as standardized or tailored tests.

The limitations of $c_9$ make it no more attractive to a test evaluator than KR20. While KR20 demands acceptance of certain assumptions, so does $c_9$, specifically that guessing is not present (or with a likelihood of less than .2) and that the Guttman hierarchy is an acceptable standard for comparison.

**Highest Covariance Method**

Another alternative to the use of the average $\phi$ was proposed by Ayabe (1994). He advocated the substitution of the highest covariate in place of the average as an estimate of intra-item reliability. He believed that the correlation between any item and its hypothetical counterpart should be at least as good as the correlation between the real item and any other item. Evaluation of this $\phi$-alternative has been limited to a few selected data sets.

**Criteria for Comparing Estimation Methods**

In 1957, White and Saltz conducted a review of techniques used to assess a test's reproducibility, homogeneity, or internal consistency. They evaluated Guttman's Reproducibility Index, Jackson's Plus Percentage Ratio, Loevinger's Index of Homogeneity, Green's Summary Statistic Method, and the $\phi$ coefficient according to four criteria: (a) maximal value, (b) minimum value, (c) known sampling distribution, and (d) evaluation technique for
individual test items. According to White and Saltz (1957), there should be a theoretical maximum value for all tests, otherwise it is difficult to compare two tests with the same reproducibility quotient. For similar reasons, there should be a theoretical minimum for all tests. Third, techniques should allow evaluation of the null hypothesis that the quotient is not different from zero. Evaluation of the null hypothesis is not possible unless the quotient has a known sampling distribution. Finally, the technique should allow evaluation of individual item as well as the whole test. If the individual items can not be evaluated, it is difficult to improve test performance by omission or inclusion of particular items (White & Saltz, 1957).

Of the techniques evaluated, Loevinger’s index and the $\phi$ coefficient are relevant to the present discussion of reliability. White and Saltz (1957) noted that the sampling distribution of Loevinger’s index was not known. The technique also was vulnerable to situations where two subjects had identical scores, but with one subject passing a particular item and the other individual failing the item. They questioned whether such an item should contribute to the total score.

The other technique of interest was the $\phi_m$ coefficient, where each item response was compared to the total test. Individuals were divided into categories of high or low test scores and passing or failing particular items. According
to White and Saltz (1957), the maximum $\phi_a$ was 1.0 and the minimum was 0. The sampling distribution was known, allowing the null hypothesis that $\phi_a$ was greater than or equal to 0 to be tested through Chi-square or Fisher’s Exact Test. The $\phi_a$ index was not affected by extremes of item difficulty and the end result was an estimate of the relationship between the item and the overall test score. Unfortunately, there was no composite estimate to be applied to the test as a whole.

Table 1 summarizes the characteristics of estimates discussed in this investigation. The upper limit for all estimation techniques is unity. The lower limit for reliability is zero for all techniques except the split-half method and Cudeck’s $c_2$, which have a lower limit of -1.00. Negative estimates of reliability have been computed using KR20 and its extensions, such as Huck’s $r_{Hu}$ and Winer’s $r_w$, in circumstances considered inadmissible by Kuder and Richardson. Any technique relying upon the sum of off-diagonal covariances as a determiner of reliability would be susceptible to estimates of reliability less than the theoretical lower limit of zero. Sampling distributions for all techniques are unknown with the exception of KR20. An approximation for the sampling distribution of KR20 for tests of 40 to 500 items was developed by Feldt (1965). Only Loevinger (1947) provided a technique for determining the homogeneity of an item with the test.
Table 1.
Summary of Proposed Techniques for the Estimation of Internal Consistency Reliability.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Proponent</th>
<th>Upper Limit</th>
<th>Lower Limit</th>
<th>Sampling Distribution</th>
<th>Item Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_A$</td>
<td>Ayabe (1994)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$r_{oo}$</td>
<td>Brown (1910), Spearman (1910)</td>
<td>1</td>
<td>-1</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>MSPLIT</td>
<td>Callender &amp; Osburn (1979)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>Cliff (1984)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$c_2$</td>
<td>Cudeck (1980)</td>
<td>1</td>
<td>-1</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$L$</td>
<td>Guttman (1945)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$r_H$</td>
<td>Horst (1953)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>Huck (1978)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>KR20</td>
<td>Kuder &amp; Richardson (1937)</td>
<td>1</td>
<td>0</td>
<td>estimated</td>
<td>no</td>
</tr>
<tr>
<td>Symbol</td>
<td>Proponent</td>
<td>Upper Limit</td>
<td>Lower Limit</td>
<td>Sampling Distribution</td>
<td>Item Statistic</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-----------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>$H_t$</td>
<td>Loevinger (1947)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>$r_R$</td>
<td>Raju (1982)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>ten Berge &amp; Zegers (1978)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>$r_w$</td>
<td>Winer (1971)</td>
<td>1</td>
<td>0</td>
<td>unknown</td>
<td>no</td>
</tr>
</tbody>
</table>
Criteria for evaluating reliability estimates were also proposed by Ayabe (1994). He suggested evaluating reliability estimates with three criteria: (a) the value should never exceed 1.0 since the true variance can never exceed the observed variance, (b) the estimate should never be negative because negative variances are not possible by definition, and (c) values should be exactly 1.0 when items are perfectly consistent. These criteria are compatible with the criteria proposed by White and Saltz (1957).

To evaluate select reliability estimation measures, Ayabe constructed four 6-by-6 item-person score matrices "to exploit the weakness of KR20" (p. 2). The selected estimation procedures were KR20, odd-even split-half method, Horst's estimate ($r_H$), average inter-item correlation (average $r$), average $\phi/\phi_{max}$, and maximum split-half (Guttman's $L_4$).

Ayabe's first pattern contained items of equal difficulty, the only condition where KR20 can equal unity. Individuals answered all items either correctly or incorrectly. This pattern is illustrated in Figure 1a. All estimation methods yielded perfect reliability estimates except for Horst's method, yielding $r_H = .952$. Ayabe observed that since Horst's method adjusted items for difficulty by rank ordering, the ordering was confounded by tied ranks. All other reliability estimates performed as
**Figure 1.**

Ayabe (1994) Test Item-by-Person Score Matrices.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

anticipated. Ayabe concluded that only Horst's method did not meet the criterion of 1.0 as the maximum reliability coefficient.

Ayabe's second pattern, a Guttman scalogram (Figure 1b), is considered to be perfectly internally consistent. The scalogram's perfection should yield reliability estimates of 1.0; however, only $r_H$, $\phi/\phi_{\text{max}}$, and maximum split-half resulted in the unity estimate. The remaining estimates were .80 by KR20, .81 by average $r$, and .92 by odd-even split-half.

The third pattern (Figure 1c) was contrived to evaluate the potential for estimates exceeding the $|1|$ boundary. Neither KR20 or $r_H$ could be calculated because computation required division by 0. Average $r$ and maximum split-half yielded estimates of 0, but $\phi/\phi_{\text{max}}$ estimated reliability as 1.0. The computation of odd-even split-half was complicated by 0; however, the addition of two more items in a consistent pattern would have provided a reliability estimate of 1.0.

Ayabe's final pattern (Figure 1d) was designed to produce negative reliability coefficients while producing an odd-even split-half reliability of unity. Negative reliabilities were computed using KR20 and $r_H$. The average $r$ was .294. Odd-even and maximum split-half estimates were 1.0. Requiring division by 0, $\phi/\phi_{\text{max}}$ could not be determined.
Ayabe concluded that KR20 and Horst's method failed all three axioms. At least one axiom was failed by every method except for maximum split-half. He acknowledged that some individuals would argue that Pattern D was not appropriate because it violated KR20's assumption of unidimensionality, but he pointed out that only Pattern A strictly met all of Kuder and Richardson's assumptions. Ayabe (1994) recommended more experimentation with "finer unbiased variation in differential item-by-person matrices" (p. 5) and suggested that measurement researchers recognize circumstances where the KR20 estimate would be inappropriate.

The evaluation criteria of White and Saltz (1957) and Ayabe (1994) accentuate the problem of evaluating the various reliability estimation methods presented in this review. There are no defined criteria for evaluation. At best, proponents of one method have presented estimates in various situations, or they have made limited comparisons to other methods. The criteria of White and Saltz (1957) and Ayabe (1994) provide a framework for comparing methods. To develop a framework, the White and Saltz (1957) and Ayabe (1994) criteria need to be inspected more closely.

There is some common ground in the two methods. Each requires a theoretical maximum value for a reliability estimate. The value for both sets of standards is 1.0. Each requires a theoretical minimum. The minimum
reliability is 0.0. Although Ayabe's (1994) criteria are satisfied at this point, White and Saltz (1957) additionally recommended knowledge of the sampling distribution of the coefficients. Knowledge of the sampling distributions may provide insight into some of the reliability estimates' non-unity non-zero results from the Ayabe patterns.

There are two distinctions between the White and Saltz and Ayabe criteria. The first distinction was the requirement for evaluation of individual items by White and Saltz, but not Ayabe.

White and Saltz (1957) also focused on reproducibility with an implied order for item difficulties, as demonstrated by the Guttman scalogram. Given a set of individual test item responses, a Guttman scalogram can be generated by arranging the subjects in a hierarchy in order of test score and item difficulties, as illustrated in Figure 1b where the individual scores were observed to decrease from top to bottom of the table and item difficulties increase from left to right. Manipulating Ayabe patterns C and D to establish a hierarchy provides additional insight. Figure 2e illustrates the rearranged Pattern D. If Ayabe's Pattern D were arranged in the recommended hierarchy, the pattern's appearance changes. It becomes clear how the pattern yields a perfect odd-even coefficient. There are identical response patterns for matching pairs of items. Other arrangements of items would yield differing coefficients.
Figure 2.
Rearranged Ayabe Test Item-by-Person Score Matrices C and D.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6</td>
<td>1 3 5 2 4 6</td>
</tr>
<tr>
<td>P</td>
<td>1 1 1 1 1 1 1</td>
<td>1 1 1 1 0 0 0</td>
</tr>
<tr>
<td>E</td>
<td>1 1 1 1 0 0 0</td>
<td>3 1 1 1 0 0 0</td>
</tr>
<tr>
<td>R</td>
<td>1 1 1 1 0 0 0</td>
<td>5 1 1 1 0 0 0</td>
</tr>
<tr>
<td>S</td>
<td>1 1 1 1 0 0 0</td>
<td>2 0 0 0 1 1 1</td>
</tr>
<tr>
<td>O</td>
<td>1 1 0 0 1 1 1</td>
<td>4 0 0 0 1 1 1</td>
</tr>
<tr>
<td>N</td>
<td>1 1 0 0 1 1 1</td>
<td>6 0 0 0 1 1 1</td>
</tr>
</tbody>
</table>

Note. From "Rederiving KR20: Reply to Kuder" by H. I. Ayabe, 1994, Unpublished manuscript. Adapted with permission.
The negative KR20 and Horst reliability estimates result from the influence of negatively discriminating items. These items might suggest the emergence of a second factor in the data, which would be a violation of the fundamental Kuder and Richardson assumption of unidimensionality.

Perhaps Ayabe's checkerboard pattern (Figure 1c) is the most interesting. Cursory inspection of the pattern might lead one to conclude that this pattern reflects a lack of reliability (Ayabe, 1994). Arranging the items and subjects to group similar responses, a very stable pattern appears, as demonstrated in Figure 2f. Visual examination of the reconfigured pattern could suggest the presence of two factors, although it can not be demonstrated mathematically. Factor analysis shows one factor. The pattern's symmetry contributes to justification of perfect reliability. While White and Saltz (1957) suggested that selection of a method for the estimation of reproducibility is of minor consequence since they are algebraically similar, this pattern reinforces the position that the particular method used for the estimation of reliability is crucial.

One distinction can be made between pattern B and the other three Ayabe patterns. In Pattern B each subject and item contributes unique information to the estimation of reliability. The other three patterns can be 'simplified.' Pattern A reflects the situation Loevinger (1947) described as "absurd" (p. 11). Items were all of equal difficulty and
there was no error, consequently each subject answers all the items correctly or incorrectly, depending on the individual’s ability compared to the item difficulty. As Loevinger noted, such a test could be reduced to one item. Pattern C was also amenable to simplification. A correct answer to item 1 implied correct answering of items 3 and 5, plus incorrect answers to items 2, 4, and 6. Pattern C is a single factor test not unlike Pattern A. All test information was contained in one test item. There is very little information obtainable from this test pattern.

Pattern B was different from Patterns A and C. Knowing an individual’s response to one particular item did not lead to knowing the rest of the matrix. Knowledge of the total score was required before the rest of the pattern could be reconstituted for that individual. The pattern meets the classic definition of homogeneity.

In Pattern D some items were redundant, reducing the test to three items, either the odd or even items. As the number of items decreased, the available amount of information was reduced. Some variance was maintained because of the difference in response patterns for items 3/4 and 5/6, but because of negatively discriminating items, computational difficulties arise. Whether these two items reflected a second factor or error is left to the interpretation of the investigator. How to calculate a reliability estimate to fit the defined range from 0 to 1 is
also left to the test designer. If it is believed that the test is unidimensional, then deletion of non-conforming items might be considered.

Krus and Helmstadter (1993) developed an alternative method for the estimation of KR20 reliability when negative coefficients are computed from standard estimation methods. Krus and Helmstadter suggested that the reliability coefficient may be negative when the variance contribution of the second principal component is the best estimate of reliability and the first principal component reflects the error contribution. Other multi-dimensional solutions have been proposed by Armor (1974) and Conger and Lipshitz (cited in Feldt & Brennan, 1988).

Summary

A review of literature concerning estimation techniques for internal consistency reliability reveals a series of investigations and essays spanning almost 60 years. Comparative research is summarized in Table 2. Of the nearly two dozen techniques described in the literature, only seven have been subjects of investigations. Experimental designs have varied, but can be generally described as non-randomized descriptive studies with small data sets. Results have been contradictory or inconclusive. There has been no investigation comparing more than five estimation techniques.
Table 2.


<table>
<thead>
<tr>
<th>Investigator</th>
<th>Comparison</th>
<th>Design</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brogden (1946)</td>
<td>KR20</td>
<td>Controlled $r_{xi}$, then varied item difficulty</td>
<td>KR20 not seriously affected</td>
</tr>
<tr>
<td>Callender &amp; Osburn (1977)</td>
<td>MSPLIT, KR20, $r_\infty$</td>
<td>9, 10-item tests and 3, 40-item tests from pools of 100 total items and samples of 100 from pool of 380 subjects</td>
<td>MSPLIT accurately estimated largest split half, always higher than KR20</td>
</tr>
<tr>
<td>Cliff (1984)</td>
<td>$r_\gamma$, KR20, $r_{xx}$, $r_{xx'}$, $r_{x'}$</td>
<td>computer-generated tests of 5 and 10 items, 100 and 500 subjects</td>
<td>effect of using $r_\gamma$ is small unless number of items is small or variance of item difficulties is large</td>
</tr>
</tbody>
</table>
Table 2 (continued).


<table>
<thead>
<tr>
<th>Investigator</th>
<th>Comparison</th>
<th>Design</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach (1951)</td>
<td>$\alpha$, $\phi$, $\phi/\phi_{\text{max}}$</td>
<td>Controlled $r_{\text{int}}$, then varied item difficulty</td>
<td>effect on $\alpha$ not important</td>
</tr>
<tr>
<td>Cronbach &amp; Azuma (1962)</td>
<td>$\rho_\alpha$, $\rho_{\text{IM2}}$, $\alpha_s$, $\alpha_L$, $r_H$, $KR_{21}$</td>
<td>range of item difficulty from .01 to .99 and varied $r_{ij}$</td>
<td>$\alpha$ is satisfactory</td>
</tr>
<tr>
<td>Cudeck (1980)</td>
<td>$KR_{20}$, $H_r$, $c_0$</td>
<td>varied discrimination, difficulty, and ability of 20- and 40-item tests with 100 subjects</td>
<td>$H_r$ in range of .3 to .5 corresponds to acceptable KR20 reliability plus provides information on consistency</td>
</tr>
</tbody>
</table>
Table 2 (continued).


<table>
<thead>
<tr>
<th>Investigator</th>
<th>Comparison</th>
<th>Design</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lord (1955)</td>
<td>KR20</td>
<td>one test of 28 items and 136 subjects</td>
<td>KR20 approaches least upper bound as established by matched-form coefficient</td>
</tr>
<tr>
<td>Raju (1982)</td>
<td>$L, r_R, r_H$</td>
<td>one matrix of 8 items and 10 subjects with difficulties from 0.3 to 1.0</td>
<td>not all properties of index are known</td>
</tr>
<tr>
<td></td>
<td>$H, KR20$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 (continued).


<table>
<thead>
<tr>
<th>Investigator</th>
<th>Comparison</th>
<th>Design</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ten Berge &amp; Zegers (1978)</td>
<td>KR20, $L_2$, $L_4$, 8 tests of 3 and 4 items</td>
<td>calculation of $L_2$ worthwhile, $\mu_2$ for tests with wide range of inter-item covariance</td>
<td></td>
</tr>
<tr>
<td>Terwilliger &amp; Lele (1979)</td>
<td>KR20, $L_2$, $L_4$, $\mu$</td>
<td>Controlled $S_x^2$, varied item difficulty and score</td>
<td>$L_4$ better estimate than $L_2$ or KR20</td>
</tr>
<tr>
<td>Woodhouse &amp; Jackson (1977)</td>
<td>KR20, $L_2$, $L_4$, glb</td>
<td>3, 3 item tests and 1 test consisting of 20 subtests</td>
<td>$L_4$ better estimate than $L_2$ or KR20</td>
</tr>
</tbody>
</table>
Criteria for evaluation of techniques are not defined either. Ayabe (1994) and White and Saltz (1957) provided frameworks with limited comparisons of estimation techniques.

The current investigation was designed to address limitations of the literature review. Eleven techniques for the estimation of internal consistency reliability or homogeneity were compared to each other and using standards derived from common criteria of Ayabe (1994) and White and Saltz (1957). Lord and Novick's (1968) concept of tau-equivalence has been included as an additional requirement for monotonic progression of reliability estimates between their upper and lower limits.
CHAPTER 3

Methodology

A series of item-by-person data matrices were constructed to compare methods for the estimation of internal consistency reliability. Each matrix represented the simulated responses of eight individuals to a test comprised of eight dichotomously-scored items assumed to be unidimensional in nature. Internal consistency reliability was estimated using the methods described by Ayabe (1994), Cliff (1984), Guttman (1945), Horst (1953), Huck (1978), Kuder and Richardson (1937), Loevinger (1947), Raju (1982), ten Berge and Zegers (1978), and Winer (1971). Reliability estimates from each method were compared to each other and preset criteria.

Sample

The sample consisted of 122 eight-by-eight item-by-person data matrices. Row data represented the simulated dichotomous responses of an individual to each of eight test items (columns).

Overview of Sample

The data matrices were obtained by making random substitutions in data sets representing perfect reliability using either Guttman or Kuder Richardson criteria (Figure 3). The starting Kuder-Richardson matrix (Figure 3a) represents the circumstance where a KR20-type reliability is 1.0. In Figure 3a, four individuals answer all of the items
correctly, while the remaining four individuals answer none of the items correctly. The starting Guttman scalogram (Figure 3b) represents the situation where Loevinger-type homogeneity is 1.0. In Figure 3b, one individual answers all the items correctly, the next individual correctly answers all but the one most difficult item, and so forth until the eighth person correctly answers only the easiest item.

Figure 3.
Perfect Item-by-Person Score Matrices.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 1 1 1</td>
<td>1 1 1 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1 1</td>
<td>1 1 1 1 1 1</td>
</tr>
<tr>
<td>3</td>
<td>1 1 1 1 1</td>
<td>1 1 1 1 1 0</td>
</tr>
<tr>
<td>4</td>
<td>1 1 1 1 1 1</td>
<td>1 1 1 1 0 0</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 0 0 0 0</td>
<td>1 1 1 1 1 0 0</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 0 0 0 0</td>
<td>1 1 1 0 0 0 0</td>
</tr>
<tr>
<td>7</td>
<td>1 1 0 0 0 0 0 0</td>
<td>1 1 1 0 0 0 0</td>
</tr>
<tr>
<td>8</td>
<td>0 0 0 0 0 0 0 0</td>
<td>1 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>
To simulate error, random substitutions were made in 120 matrices. There were nine clusters of 10 matrices each, representing substitutions in two, four, and six rows from the original Guttman matrix; two, four, and six columns from the original Guttman matrix; and two, four, and six columns from the original Kuder-Richardson matrix. For each cluster, the specific rows or columns to be substituted were identified from a series of random numbers generated for each data matrix using the random number function of Excel® spreadsheet for Windows® (Microsoft Corporation, Redmond, WA). Each of the substitutions was made with a randomly generated pattern meeting the same row or column criteria. For example, if the row selected for random substitution had six correct followed by two incorrect answers, then the new row also contained exactly six correct answers but the order for these correct answers was determined at random.

Thirty additional data matrices were generated completely at random. The sole criterion for item and subject distribution was that exactly half the items be answered correctly. The final two data matrices were the perfect Guttman scale and the perfect Kuder-Richardson tests.

Representativeness of Sample

To assure randomness of selection, the distribution of rows and columns selected for substitution was evaluated using the Chi-square test for one sample. The frequency
distribution for the selection of a particular row or column for random substitution is presented in Table 3. The variation in number of substitutions was not statistically significant ($\chi^2 = 4.76, df = 6, p > .50$).

Table 3.
Frequency Distribution of Row and Column Numbers Selected for Substitution.

<table>
<thead>
<tr>
<th>Row/Column Number</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
</tr>
</tbody>
</table>

Once randomness was confirmed, the data sets were screened for matrices with negative average off-diagonal intercorrelations, in accordance with Kuder and Richardson's (1937) belief that reliability was a function of positive intercorrelation. Twenty-four (19.7%) data matrices were excluded from the investigation, including 7 Guttman matrices with column substitution, 4 Kuder-Richardson
matrices, and 13 random data matrices. The final sample for subsequent analysis consisted of 98 matrices.

Adequacy of the distribution of item and test characteristics was assessed by inspection of the frequency distributions of item variance, standard deviation of item difficulty and KR20 reliability estimates. Because of the selection criteria for data matrices, it was anticipated that the distribution of sample characteristics would be representative but not be uniform.

Mean item variances ranged from .164 to .250 (Table 4). The maximum upper limit for item variance is .250 which occurs when half the test takers respond correctly to the item, such as in the perfect KR20-type item-by-person matrix (Figure 3a). Smaller mean item variances indicate a wider range of item difficulties.

The standard deviation of item difficulties ranged from .000 to .286 (Table 5). Lack of variance is characteristic of the KR20-type patterns where each question was answered correctly by half the test takers. A broader range of item difficulties is reflected by larger standard deviations of item difficulties.

Manipulation of the data matrices yielded a wide range of reliability estimates. Kuder-Richardson formula 20 estimates ranged from .041 to 1.00 (Table 6). Based on the wide distribution of item variances, standard deviation of item difficulties, and KR20 reliability estimates, the data
sets were judged to be representative of a wide range of situations that may be encountered in the development of unidimensional tests.

Table 4.
Frequency Distribution of Mean Item Variance by Item-by-Person Matrix.

<table>
<thead>
<tr>
<th>Mean Item Variance</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1625 - .1874</td>
<td>30</td>
</tr>
<tr>
<td>.1875 - .2124</td>
<td>10</td>
</tr>
<tr>
<td>.2125 - .2374</td>
<td>27</td>
</tr>
<tr>
<td>≥ .2375</td>
<td>31</td>
</tr>
</tbody>
</table>

Table 5.
Frequency Distribution of Standard Deviation of Item Difficulties by Item-by-Person Matrix.

<table>
<thead>
<tr>
<th>Standard Deviation of Item Difficulty</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.000 - .009</td>
<td>27</td>
</tr>
<tr>
<td>.010 - .019</td>
<td>3</td>
</tr>
<tr>
<td>.020 - .029</td>
<td>33</td>
</tr>
<tr>
<td>.030 - .039</td>
<td>35</td>
</tr>
</tbody>
</table>
Table 6.
Frequency Distribution of Kuder-Richardson Reliability Estimates by Item-by-Person Matrix.

<table>
<thead>
<tr>
<th>KR20 Reliability Estimates</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.041 - .199</td>
<td>15</td>
</tr>
<tr>
<td>.200 - .399</td>
<td>12</td>
</tr>
<tr>
<td>.400 - .599</td>
<td>15</td>
</tr>
<tr>
<td>.600 - .799</td>
<td>41</td>
</tr>
<tr>
<td>≥ .800</td>
<td>15</td>
</tr>
</tbody>
</table>

Instrumentation

No special instrumentation was used in this investigation. Reliability estimates were computed using Excel® spreadsheet on an IBM personal computer. A master spreadsheet was constructed using the reliability estimation formulas from the proponents' original articles; except for the computation of Loevinger's $H_l$, where the computational formula proposed by Raju (1982) was used. The accuracy of each formula as entered into the computer was tested by estimating reliability from perfect Guttman and Kuder Richardson matrices. In circumstances where an author provided a computational example, the reliability of the sample data set was also computed using the same formulas as in the template.
Many techniques required the ranking of test scores. When scores were tied, each score was ranked as the arithmetical mean. For example, if the second and third highest test scores were tied, the rank of 2.5 was assigned to each.

To avoid transcription errors, once a numerical value was generated, it was not transcribed. Calculated values were linked within the individual spreadsheet as well as to a summary spreadsheet that was used for graphic analysis. Data were rounded to $10^3$ and exported to SPSS/PC® (SPSS, Inc., Chicago, IL) for additional statistical analysis.

**Methods**

Evaluation of the various reliability estimation methods required the development of a plan for the manipulation of covariance within the data set. Once a manipulation system was developed, the effect of changes in the item variances or inter-item covariances could be evaluated.

**Data Set Generation**

The original data set consisted of 122 data matrices representing the responses of eight hypothetical individuals to each of eight items on a hypothetical test. The matrix patterns were developed using preset criteria.

The starting point for 60 of the matrices was a Guttman-type data matrix described as possessing perfect reliability. For 10 matrices each; two, four, or six rows
or columns to be substituted were randomly selected by computer using the random number generating function of Excel®. For each matrix a series of random numbers between two and eight were generated to identify which of the rows or columns would be manipulated. Number one was not included because the first row and column consisted of only correct responses and could not be manipulated. Consecutive unique numbers were used to identify which columns or rows would be substituted in each matrix. A new set of random numbers was generated for each matrix.

The replacement rows and columns were generated randomly using Excel®'s Bernoulli distribution function. Theoretical probability levels for success with each trial were assigned to achieve the desired mix of correct and incorrect responses and a series of eight digit patterns were derived. Since the theoretical probability of success did not always coincide with the identical outcome over the eight digits, the generated output was sorted by actual probabilities before assignment to a particular data matrix.

Starting from the Guttman distribution, row substitution resulted in manipulation of the covariance by altering the item variances while controlling the overall test score variance. Column substitution manipulated covariance by controlling item variance and allowing overall test score variance to vary. The strategy of substitution based on rows and columns provided the opportunity to alter
either item variance or total test variance in a data set without changing the second variance parameter. Increasing the number of rows or columns substituted was intended to increase the amount of error introduced into the data matrix and provide a progressive decrease in reliability.

Because the Kuder-Richardson pattern consisted of rows where all responses were either correct or incorrect, it was not possible to perform row substitution. Column substitution was conducted as described for the Guttman patterns in order to evaluate the effect of manipulating total test score variance with a fixed item variance.

The complete data set also included thirty data matrices that were randomly generated using the Bernoulli random number generation function of Excel®. Matrices were generated with the only preset criteria that exactly half the items be answered correctly. When the actual output of the random sequence did not meet this criteria, the data matrix was discarded. The generation of random matrixes was continued until 30 data matrices were obtained.

**Design**

A descriptive design was used to evaluate estimates of reliability. Data matrices were constructed to manipulate the variance-covariance structure through the progressive introduction of error. Reliability estimates were computed for each matrix.
Data matrices were constructed to exploit the weakness of reliability estimates. Since estimates of reliability tend to increase as the number of items increase (Cronbach, 1951; Sax, 1989), data sets were intentionally restricted to eight items. Smaller test sizes also tend to increase the disparity between estimates calculated with different methods.

Distribution of test scores affects reliability (Brogden, 1946). A rectilinear arrangement of test scores was chosen as the starting point for this investigation based on Brogden's demonstration that the difference between KR20 and KR2 reliability estimates was greatest with the rectilinear arrangement.

Reliability estimates were compared with each other. It was assumed that indexes measuring the same construct (internal consistency reliability) should yield similar results. Results also were compared to Guttman's $L_1$ which is the ratio of off-diagonal covariance to total score variance. Guttman's $L_1$ was chosen because it reflects a fundamental concept that reliability is a function of the inter-relatedness of items (Callender & Osburn, 1977; Kuder & Richardson, 1937). By virtue of the conservative nature of the off-diagonal covariance/total variance ratio, $L_1$ serves as a lower bound for reliability estimates.
Threats to Validity

The investigation was reviewed for the influence of the 12 threats to validity described by Campbell and Stanley (1963). Internal validity was not influenced by event history, maturation, effect of testing, instrumentation, selection, or selection-maturation interaction. There were two threats to internal validity. The sample selection process for this investigation was vulnerable to loss of data matrices (experimental mortality) possessing potentially low reliability because of negative off-diagonal covariances. The disproportionate loss of data sets from specific clusters in an otherwise balanced sample contributed to statistical regression favoring matrices with potentially higher reliability coefficients. Although balanced representation of data matrices from across the entire span of potential reliability estimates was desirable, the loss of low reliability data matrices was not considered critical since the use of tests with low or negative reliability estimates should be avoided.

The principal threat to external validity was the selection-experimental variable interaction. The emphasis on covariance as the primary contributor to test reliability favored the performance of estimation methods derived directly from the concept of the off-diagonal covariance to total test variance ratio. The three remaining threats to external validity: reactive effect of testing, reactive
effect of testing arrangements, and multiple-treatment interference were not applicable to this investigation.

**Statistical Analysis**

Reliability estimation techniques were evaluated in relation to pre-established criteria. Mandatory requirements for an adequate estimator of reliability were (a) an upper limit of 1.00, (b) a lower limit of .00, (c) a monotonic progression of values between the upper and lower limits. An additional comparative requirement was association among reliability estimates.

Evaluation of the upper limit of reliability was conducted by examination of the frequency distributions for each reliability estimator. There should be no circumstance where a reliability estimate exceeded 1.00. An estimate of exactly 1.00 should be computed given a "perfect" pattern (either Guttman or Kuder-Richardson).

Simple linear regression was used to project the lower limit of each reliability estimator. The Y-intercept for linear regression of the reliability estimator should be 0.000.

Monotonic progression of estimates between the upper and lower limits was explored using regression. Goodness of Fit and linearity were evaluated using $R^2$ and ANOVA.

Indexes of association among reliability estimation techniques were computed as Pearson product-moment correlation coefficients. Analysis was conducted for the
entire data set as well as subsets consisting of all Guttman matrices, Guttman matrices with row substitution, Guttman matrices with column substitution, Kuder-Richardson matrices with column substitution, and random matrices.

The magnitude of difference between estimation techniques was computed as the difference between the estimate and KR20. Kuder-Richardson formula 20 was used as an index because most estimation procedures introduced subsequent to KR20 were proposed as improvements on the Kuder-Richardson method.

Limitations and Delimitations

Factors affecting test reliability included the number of test items and shape of the distribution of test scores. This investigation used only tests with eight items. Performance of reliability estimators may vary with larger or smaller numbers of test items.

This investigation included only estimation techniques that could be easily computed by a classroom teacher without sophisticated computer resources or skills. Estimation techniques that required computer searching, such as greatest lower bound (Woodhouse & Jackson, 1977), or individualized algorithms, such as MSPLIT (Callender & Osburn, 1977) were not included. Methods that did not provide a discrete estimate, such as Guttman’s $L_4$, were not included, nor were techniques with limited application, such as Guttman’s $L_5$ and $L_6$. 
The most difficult constraint to evaluation of reliability estimates was the attempt to quantify a latent trait. It was assumed that each data matrix represented test scores intended to quantify one latent trait. In actual testing situations, unidimensionality should be validated using methods described by Hattie (1984).

Caution must be used when generalizing the results of this investigation. Although a wide range of data matrices have been tested, this sample is, at best, representative of a restricted range of possible testing situations.
CHAPTER 4
Findings

Statistical analysis was conducted to explore the limits of reliability estimation techniques, the linearity of progression of estimates between the upper and lower limits, and the relationship among estimates.

General Sample Characteristics

Separate analyses were conducted for the overall sample as well as the Guttman pattern with row substitution, Guttman pattern with column substitution, Kuder-Richardson, and random clusters. Characteristics of the overall sample are presented in Table 7. Total variance ($S^2$), item variance ($pq$), standard deviation of item difficulty ($S_p$), and total score ($X$) were functions of the design of the computer-generated sample. Low minimum and high maximum values for the reliability estimation techniques reflected the overall range of test patterns in the sample. Loevinger’s technique, $H$, yielded the lowest overall estimate of reliability, .323. The highest value, .845, was estimated using the $r_A$ technique described by Ayabe (1994).

Table 8 summarizes the characteristics of the 30 Guttman patterns with random row substitutions and the perfect Guttman pattern. In this cluster of the sample, covariance was manipulated varying item difficulty while total variance remained the same for each matrix. The designed control of test scores was reflected in the absence
Table 7.
Overall Sample Characteristics ($n = 98$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>$S$</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_x^2$</td>
<td>4.008</td>
<td>1.797</td>
<td>0.750</td>
<td>9.500</td>
</tr>
<tr>
<td>$\Sigma p q$</td>
<td>1.706</td>
<td>0.268</td>
<td>1.312</td>
<td>2.000</td>
</tr>
<tr>
<td>$S_p$</td>
<td>0.018</td>
<td>0.030</td>
<td>0.000</td>
<td>0.286</td>
</tr>
<tr>
<td>$X$</td>
<td>4.276</td>
<td>0.250</td>
<td>4.000</td>
<td>4.500</td>
</tr>
<tr>
<td>$H_i$</td>
<td>0.323</td>
<td>0.236</td>
<td>0.009</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_i$</td>
<td>0.512</td>
<td>0.222</td>
<td>0.036</td>
<td>0.875</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>0.567</td>
<td>0.269</td>
<td>-0.021</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_w$</td>
<td>0.567</td>
<td>0.265</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>KR20</td>
<td>0.585</td>
<td>0.254</td>
<td>0.041</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_y$</td>
<td>0.586</td>
<td>0.256</td>
<td>0.040</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_H$</td>
<td>0.625</td>
<td>0.270</td>
<td>0.044</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_R$</td>
<td>0.701</td>
<td>0.270</td>
<td>0.061</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.729</td>
<td>0.161</td>
<td>0.364</td>
<td>1.000</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.737</td>
<td>0.157</td>
<td>0.376</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.845</td>
<td>0.106</td>
<td>0.521</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 8.
Sample Characteristics For Guttman Scales with Row Substitutions (n = 31).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^2_x$</td>
<td>5.250</td>
<td>0.000</td>
<td>5.250</td>
<td>5.250</td>
</tr>
<tr>
<td>$\Sigma pq$</td>
<td>1.682</td>
<td>0.160</td>
<td>1.312</td>
<td>1.906</td>
</tr>
<tr>
<td>$S_p$</td>
<td>0.027</td>
<td>0.048</td>
<td>0.009</td>
<td>0.086</td>
</tr>
<tr>
<td>$X$</td>
<td>4.500</td>
<td>0.000</td>
<td>4.500</td>
<td>4.500</td>
</tr>
<tr>
<td>$H_l$</td>
<td>0.502</td>
<td>0.167</td>
<td>0.301</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_1$</td>
<td>0.680</td>
<td>0.030</td>
<td>0.637</td>
<td>0.750</td>
</tr>
<tr>
<td>$r_w$</td>
<td>0.767</td>
<td>0.036</td>
<td>0.716</td>
<td>0.851</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>0.771</td>
<td>0.042</td>
<td>0.713</td>
<td>0.877</td>
</tr>
<tr>
<td>KR20</td>
<td>0.777</td>
<td>0.035</td>
<td>0.728</td>
<td>0.857</td>
</tr>
<tr>
<td>$r_y$</td>
<td>0.787</td>
<td>0.061</td>
<td>0.713</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.826</td>
<td>0.026</td>
<td>0.775</td>
<td>0.890</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.832</td>
<td>0.026</td>
<td>0.778</td>
<td>0.895</td>
</tr>
<tr>
<td>$r_H$</td>
<td>0.836</td>
<td>0.063</td>
<td>0.746</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_R$</td>
<td>0.906</td>
<td>0.040</td>
<td>0.849</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.915</td>
<td>0.024</td>
<td>0.842</td>
<td>0.955</td>
</tr>
</tbody>
</table>
of variance in test scores. The perfect Guttman pattern was included to assure a maximum upper limit to reliability. Lower limits of reliability were higher than the overall sample. The lowest mean reliability was .502, computed using Loevinger's $H_t$ technique. Ayabe's $r_A$ produced the highest mean reliability, .915.

The effects of random column substitution in the Guttman scale are summarized in Table 9. Absence of item variance was the result of controlling item difficulties while allowing test variance to vary. An upper limit of 1.00 for reliability was established through the inclusion of the perfect Guttman scale with the other 23 data matrices used for analysis. The lower limits of reliability were lower overall than those obtained from the row substitution cluster. The lower limits for $L_2$, $\mu_2$, and $r_A$ represented the overall lower limits for these three estimation techniques. Overall mean estimates of reliability were lower than the Guttman patterns with row substitutions. The lowest mean reliability estimate, .349, was computed using Loevinger's $H_t$. The highest mean reliability estimate, .786, was computed with Ayabe's $r_A$.

Characteristics of the Kuder-Richardson cluster with column substitution are presented in Table 10. Variation in item difficulty was constrained by the Kuder-Richardson model where all items were answered either correctly or incorrectly. The maximum upper limit for reliability was
Table 9.
Sample Characteristics For Guttman Scales with Column Substitutions (n = 24).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_x^2$</td>
<td>2.688</td>
<td>0.987</td>
<td>1.500</td>
<td>5.250</td>
</tr>
<tr>
<td>$\Sigma pq$</td>
<td>1.313</td>
<td>0.000</td>
<td>1.312</td>
<td>1.312</td>
</tr>
<tr>
<td>$S_p$</td>
<td>0.039</td>
<td>0.053</td>
<td>0.029</td>
<td>0.286</td>
</tr>
<tr>
<td>$X$</td>
<td>4.500</td>
<td>0.000</td>
<td>4.500</td>
<td>4.500</td>
</tr>
<tr>
<td>$H_1$</td>
<td>0.349</td>
<td>0.251</td>
<td>0.048</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_1$</td>
<td>0.445</td>
<td>0.201</td>
<td>0.125</td>
<td>0.750</td>
</tr>
<tr>
<td>$r_w$</td>
<td>0.487</td>
<td>0.240</td>
<td>0.106</td>
<td>0.851</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>0.501</td>
<td>0.241</td>
<td>0.089</td>
<td>0.877</td>
</tr>
<tr>
<td>KR20</td>
<td>0.508</td>
<td>0.230</td>
<td>0.143</td>
<td>0.857</td>
</tr>
<tr>
<td>$r_r$</td>
<td>0.530</td>
<td>0.250</td>
<td>0.146</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_H$</td>
<td>0.593</td>
<td>0.268</td>
<td>0.167</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_R$</td>
<td>0.660</td>
<td>0.259</td>
<td>0.222</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.660</td>
<td>0.150</td>
<td>0.364</td>
<td>0.890</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.669</td>
<td>0.149</td>
<td>0.376</td>
<td>0.895</td>
</tr>
<tr>
<td>$r_\lambda$</td>
<td>0.786</td>
<td>0.118</td>
<td>0.521</td>
<td>0.952</td>
</tr>
</tbody>
</table>
Table 10.
Sample Characteristics For KR Scales ($n = 27$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>$S$</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{2}^2$</td>
<td>4.740</td>
<td>2.359</td>
<td>0.750</td>
<td>9.500</td>
</tr>
<tr>
<td>$\Sigma pq$</td>
<td>2.000</td>
<td>0.000</td>
<td>2.000</td>
<td>2.000</td>
</tr>
<tr>
<td>$S_p$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$X$</td>
<td>4.000</td>
<td>0.000</td>
<td>4.000</td>
<td>4.000</td>
</tr>
<tr>
<td>$H_t$</td>
<td>0.275</td>
<td>0.235</td>
<td>0.018</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_1$</td>
<td>0.553</td>
<td>0.226</td>
<td>0.111</td>
<td>0.875</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>0.608</td>
<td>0.275</td>
<td>0.071</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_{y}$</td>
<td>0.613</td>
<td>0.250</td>
<td>0.123</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_{w}$</td>
<td>0.616</td>
<td>0.269</td>
<td>0.090</td>
<td>1.000</td>
</tr>
<tr>
<td>KR20</td>
<td>0.632</td>
<td>0.258</td>
<td>0.127</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_{H}$</td>
<td>0.632</td>
<td>0.258</td>
<td>0.127</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_{R}$</td>
<td>0.722</td>
<td>0.255</td>
<td>0.174</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.801</td>
<td>0.164</td>
<td>0.436</td>
<td>1.000</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.808</td>
<td>0.158</td>
<td>0.460</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_{A}$</td>
<td>0.885</td>
<td>0.085</td>
<td>0.708</td>
<td>1.000</td>
</tr>
</tbody>
</table>
assured by including the perfect Kuder-Richardson data matrix. Lower limits of reliability approached, but did not meet, the lower limits of the overall data set. The lowest mean reliability estimate, .275, was computed using Loevinger's $H_t$. The highest mean estimate, .885, was computed using the Ayabe's $r_A$ method.

The characteristics of data matrices with random substitution are presented in Table 11. Overall test variance, item variance, and item difficulty were allowed to vary. The only parameter controlled by the experimental design was each individual's test score. It was anticipated the cluster of random responses would reflect lower limits of reliability. The lowest reliability estimates computed using the Loevinger's $H_t$, Huck's $r_{Hu}$, Guttman's $L_t$, Winer's $r_w$, Cliff's $r_\gamma$, KR20, Horst's $r_H$, and Raju's $r_R$ were from this cluster of random responses. The lowest mean reliability estimate, .076, was computed using $H_t$. The highest mean was .743 computed with $r_A$.

Upper Limits of Reliability

To determine the upper limit for each reliability estimation technique, the frequency distribution of estimates for each method was examined. Table 12 summarizes the maximum reliability estimate computed from the Guttman and Kuder-Richardson clusters. Maximum reliabilities were attained only in the perfect cases. Cliff's $r_\gamma$, Horst's $r_H$, Loevinger's $H_t$, and Raju's $r_R$ achieved the theoretical upper
Table 11.
Sample Characteristics For Random Matrices (n = 17).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>$S$</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^2$</td>
<td>2.515</td>
<td>0.596</td>
<td>1.750</td>
<td>4.000</td>
</tr>
<tr>
<td>$\Sigma pq$</td>
<td>1.814</td>
<td>0.085</td>
<td>1.656</td>
<td>1.938</td>
</tr>
<tr>
<td>$S_p$</td>
<td>0.015</td>
<td>0.004</td>
<td>0.009</td>
<td>0.021</td>
</tr>
<tr>
<td>$X$</td>
<td>4.000</td>
<td>0.000</td>
<td>4.000</td>
<td>4.000</td>
</tr>
<tr>
<td>$H_i$</td>
<td>0.076</td>
<td>0.056</td>
<td>0.009</td>
<td>0.186</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>0.239</td>
<td>0.178</td>
<td>-0.210</td>
<td>0.563</td>
</tr>
<tr>
<td>$L_1$</td>
<td>0.247</td>
<td>0.146</td>
<td>0.036</td>
<td>0.516</td>
</tr>
<tr>
<td>$r_W$</td>
<td>0.252</td>
<td>0.174</td>
<td>0.000</td>
<td>0.572</td>
</tr>
<tr>
<td>$r_{\gamma}$</td>
<td>0.279</td>
<td>0.164</td>
<td>0.040</td>
<td>0.578</td>
</tr>
<tr>
<td>KR20</td>
<td>0.283</td>
<td>0.167</td>
<td>0.041</td>
<td>0.589</td>
</tr>
<tr>
<td>$r_H$</td>
<td>0.297</td>
<td>0.174</td>
<td>0.044</td>
<td>0.606</td>
</tr>
<tr>
<td>$r_h$</td>
<td>0.372</td>
<td>0.204</td>
<td>0.061</td>
<td>0.717</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.544</td>
<td>0.101</td>
<td>0.409</td>
<td>0.741</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.555</td>
<td>0.099</td>
<td>0.424</td>
<td>0.749</td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.743</td>
<td>0.086</td>
<td>0.578</td>
<td>0.885</td>
</tr>
</tbody>
</table>
### Table 12.

**Reliability Estimation Techniques**

**Achieving Upper Limit of 1.0.**

<table>
<thead>
<tr>
<th>Estimation Technique</th>
<th>Guttman</th>
<th>Kuder-Richardson</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_N$</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$H_n$</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_R$</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_A$</td>
<td>------</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_{Ha}$</td>
<td>------</td>
<td>1.000</td>
</tr>
<tr>
<td>KR20</td>
<td>------</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_2$</td>
<td>------</td>
<td>1.000</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>------</td>
<td>1.000</td>
</tr>
<tr>
<td>$r_W$</td>
<td>------</td>
<td>1.000</td>
</tr>
<tr>
<td>$L_1$</td>
<td>------</td>
<td>------</td>
</tr>
</tbody>
</table>
limit of 1.0 for both Guttman and Kuder-Richardson-type data matrices. Guttman’s $L_1$ failed to achieve the upper limit in either data matrix. The remaining techniques achieved the theoretical maximum only in the Kuder-Richardson cluster.

Lower Limits of Reliability

The lower limit of reliability for each estimation technique was projected using linear regression. Results are presented in Table 13. It was anticipated that the $Y$-intercept of the regression equation would equal zero. The $Y$-intercepts for Horst’s $r_H$, KR20, and Cliff’s $r_7$ included zero within the 95% confidence limits. Negative $Y$-intercepts were projected for Huck’s $r_{Hh}$, Loevinger’s $H_l$, and Winer’s $r_w$. The $Y$-intercept for the Raju’s $r_R$ technique was mildly positive. Ayabe’s $r_A$, Guttman’s $L_2$, and ten Berge and Zegers’ $\mu_2$ had markedly positive $Y$-intercepts.

Linear Progression Between Limits

Goodness of fit of the reliability estimates to the general linear model was tested using $R^2$ and the $F$-ratio. Results of the testing are presented in Table 14. Goodness of fit was statistically significant ($p \leq .004$) for all reliability estimation methods in the overall sample and each of the clusters. When an $R^2$ less than .90 was reported, residual analysis was conducted by examination of the bivariate scatterplots of the relationship between the predicted and residual values for each score.
Table 13.
Estimated Lower Limits of Reliability Estimation Techniques.

<table>
<thead>
<tr>
<th>Estimation Technique</th>
<th>Y-intercept</th>
<th>p</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_t$</td>
<td>-.141</td>
<td>.001</td>
<td>-.2038, -.0783</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>-.052</td>
<td>.001</td>
<td>-.0594, -.0449</td>
</tr>
<tr>
<td>$r_w$</td>
<td>-.042</td>
<td>.001</td>
<td>-.0429, -.0425</td>
</tr>
<tr>
<td>$r_7$</td>
<td>-.002</td>
<td>.727</td>
<td>-.0141, .0089</td>
</tr>
<tr>
<td>KR20</td>
<td>.000</td>
<td>.374</td>
<td>-.0001, .0001</td>
</tr>
<tr>
<td>$r_H$</td>
<td>.009</td>
<td>.377</td>
<td>-.0108, .0285</td>
</tr>
<tr>
<td>$r_R$</td>
<td>.086</td>
<td>.001</td>
<td>.0679, .1048</td>
</tr>
<tr>
<td>$L_2$</td>
<td>.373</td>
<td>.001</td>
<td>.3508, .3948</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>.389</td>
<td>.001</td>
<td>.3665, .4107</td>
</tr>
<tr>
<td>$r_A$</td>
<td>.624</td>
<td>.001</td>
<td>.6017, .6472</td>
</tr>
<tr>
<td>$L_1$</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
</tbody>
</table>
Table 14.

$R^2$ and $F$-Ratios for Goodness of Fit and Linearity.

<table>
<thead>
<tr>
<th></th>
<th>Guttman Overall</th>
<th>Guttman Rows</th>
<th>Guttman Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F^*$</td>
<td>$R^2$</td>
<td>$F^{**}$</td>
</tr>
<tr>
<td>KR20</td>
<td>49811667</td>
<td>1.00</td>
<td>380201</td>
</tr>
<tr>
<td>$r_w$</td>
<td>37839095</td>
<td>1.00</td>
<td>337488</td>
</tr>
<tr>
<td>$r_H$</td>
<td>33968</td>
<td>.99</td>
<td>1253</td>
</tr>
<tr>
<td>$r_7$</td>
<td>11215</td>
<td>.99</td>
<td>358</td>
</tr>
<tr>
<td>$H_t$</td>
<td>255</td>
<td>.73</td>
<td>619</td>
</tr>
<tr>
<td>$r_H$</td>
<td>4565</td>
<td>.98</td>
<td>2284</td>
</tr>
<tr>
<td>$r_R$</td>
<td>5206</td>
<td>.98</td>
<td>296196</td>
</tr>
<tr>
<td>$L_1$</td>
<td>---</td>
<td>1.00</td>
<td>---</td>
</tr>
<tr>
<td>$L_2$</td>
<td>1203</td>
<td>.93</td>
<td>112</td>
</tr>
<tr>
<td>$r_A$</td>
<td>441</td>
<td>.82</td>
<td>17</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>1163</td>
<td>.92</td>
<td>89</td>
</tr>
</tbody>
</table>

* $(df = 1, 96)$

** $(df = 1, 30)$

*** $(df = 1, 23)$
Table 14 (continued).

$R^2$ and $F$-Ratios for Goodness of Fit and Linearity.

<table>
<thead>
<tr>
<th></th>
<th>Kuder-Richardson</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F'$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>KR20</td>
<td>15281457</td>
<td>1.00</td>
</tr>
<tr>
<td>$r_w$</td>
<td>8976398</td>
<td>1.00</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>14163280</td>
<td>1.00</td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>15606</td>
<td>.99</td>
</tr>
<tr>
<td>$H_1$</td>
<td>71</td>
<td>.74</td>
</tr>
<tr>
<td>$r_H$</td>
<td>15281457</td>
<td>1.00</td>
</tr>
<tr>
<td>$r_R$</td>
<td>1717</td>
<td>.99</td>
</tr>
<tr>
<td>$L_1$</td>
<td>---</td>
<td>1.00</td>
</tr>
<tr>
<td>$L_2$</td>
<td>202</td>
<td>.89</td>
</tr>
<tr>
<td>$r_\lambda$</td>
<td>88</td>
<td>.78</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>196</td>
<td>.89</td>
</tr>
</tbody>
</table>

* (df = 1, 25)

** (df = 1, 15)
Figures 4 and 5 illustrate representative residual plots. The typical scatterplot for estimates with a lower coefficient of determination \( R^2 \), such as \( r_A, L_2 \), and \( \mu \), is shown in Figure 4. The gradual narrowing of the data points suggested inequality of variance rather than lack of fit.

The scatterplot for the residual analysis of the Loewinger's \( H_i \) estimator is shown in Figure 5. The pattern is characteristic of non-linearity.

**Figure 4.**

**Residual Scatterplot for \( r_A \) and \( L_1 \).**

Standardized Scatterplot

<table>
<thead>
<tr>
<th>Out</th>
<th>*PREDICTED</th>
<th>*RESIDUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Symbols:

- Max N
- \( * \): 1.0
- \( : \): 2.0
- \( * \): 5.0
Figure 5.
Residual Scatterplot for $H_i$ and $L_i$.

Standardized Scatterplot
Across - *PREDICTED  Down - *RESIDUAL

Symbols:
Max N

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>7.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Association Among Estimation Techniques

Pearson product-moment correlation coefficients were computed between reliability estimates to indicate the extent of association among techniques. The results are reported in Tables 15 through 19. There were significant relationships among all the estimation techniques ($df = 96$, $p < .001$). Among the variables, the weakest relationships were between Loevinger's $H_i$ and the other estimates except for the Guttman cluster with row substitution and the cluster of random data sets where the weakest associations were between the Ayabe's $r_A$ and the other techniques.
Table 15.
Product-Moment Correlation Coefficients Among Reliability Estimates for Complete Data Sets ($n = 98$).

<table>
<thead>
<tr>
<th>KR20</th>
<th>$r_w$</th>
<th>$r_{Hu}$</th>
<th>$r_\gamma$</th>
<th>$H_1$</th>
<th>$r_H$</th>
<th>$r_R$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$r_A$</th>
<th>$\mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_w$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>.99</td>
<td>.99</td>
<td>.99 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_1$</td>
<td>.85</td>
<td>.85</td>
<td>.86  .89 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_H$</td>
<td>.99</td>
<td>.99</td>
<td>.99  .99 .90 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_R$</td>
<td>.99</td>
<td>.99</td>
<td>.99  .99 .85  .99 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td>1.00</td>
<td>1.00</td>
<td>.99  .99 .85  .99  .99 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_2$</td>
<td>.96</td>
<td>.96</td>
<td>.96  .95  .78  .94  .94  .96 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_A$</td>
<td>.91</td>
<td>.91</td>
<td>.90  .89  .74  .88  .89  .91  .95 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>.96</td>
<td>.96</td>
<td>.95  .78  .93  .94  .96  .99  .95 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 16.
Product-Moment Correlation Coefficients Among Reliability Estimates for Guttman Row Data Sets (n = 31).

<table>
<thead>
<tr>
<th></th>
<th>KR20</th>
<th>$r_w$</th>
<th>$r_\text{Hu}$</th>
<th>$r_\gamma$</th>
<th>$r_H$</th>
<th>$r_R$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$r_A$</th>
<th>$\mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR20</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$r_w$</td>
<td></td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
<td>0.89</td>
<td>0.61</td>
<td>0.87</td>
</tr>
<tr>
<td>$r_\text{Hu}$</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.89</td>
<td>0.60</td>
<td>0.86</td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>1.00</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
<td>0.87</td>
<td>0.58</td>
<td>0.85</td>
</tr>
<tr>
<td>$H_1$</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
<td>0.88</td>
<td>0.56</td>
<td>0.84</td>
</tr>
<tr>
<td>$r_H$</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.88</td>
<td>0.59</td>
<td>0.87</td>
</tr>
<tr>
<td>$r_R$</td>
<td>1.00</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>0.88</td>
<td>0.60</td>
<td>0.81</td>
</tr>
<tr>
<td>$L_1$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>0.89</td>
<td>0.60</td>
<td>0.86</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.89</td>
<td>0.89</td>
<td>0.87</td>
<td>0.86</td>
<td>0.88</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.61</td>
<td>0.60</td>
<td>0.58</td>
<td>0.56</td>
<td>0.59</td>
<td>0.60</td>
<td>0.81</td>
<td>0.89</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.87</td>
<td>0.87</td>
<td>0.85</td>
<td>0.84</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
<td>0.98</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Table 17.

Product-Moment Correlation Coefficients Among Reliability Estimates for Guttman Column Data Sets (n = 24).

<table>
<thead>
<tr>
<th>KR20</th>
<th>$r_w$</th>
<th>$r_{Hu}$</th>
<th>$r_\gamma$</th>
<th>$H_t$</th>
<th>$r_H$</th>
<th>$r_R$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$r_A$</th>
<th>$\mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR20</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_w$</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_t$</td>
<td>.94</td>
<td>.94</td>
<td>.94</td>
<td>.97</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_H$</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
<td>.94</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_R$</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.91</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
<td>.94</td>
<td>1.00</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_2$</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>.98</td>
<td>.92</td>
<td>.98</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$r_A$</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
<td>.88</td>
<td>.96</td>
<td>.97</td>
<td>.96</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>.99</td>
<td>.99</td>
<td>.97</td>
<td>.98</td>
<td>.92</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 18.
Product-Moment Correlation Coefficients Among Reliability Estimates for Kuder-Richardson Data Sets \((n = 27)\).

<table>
<thead>
<tr>
<th>KR20</th>
<th>(r_w)</th>
<th>(r_H)</th>
<th>(r_\gamma)</th>
<th>(H_t)</th>
<th>(r_R)</th>
<th>(L_1)</th>
<th>(L_2)</th>
<th>(r_A)</th>
<th>(\mu_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR20</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r_w)</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r_H)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r_\gamma)</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(H_t)</td>
<td>.86</td>
<td>.86</td>
<td>.86</td>
<td>.87</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r_H)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.86</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r_R)</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.80</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(L_1)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.86</td>
<td>1.00</td>
<td>.99</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>(L_2)</td>
<td>.94</td>
<td>.94</td>
<td>.94</td>
<td>.93</td>
<td>.79</td>
<td>.94</td>
<td>.93</td>
<td>.94</td>
<td>1.00</td>
</tr>
<tr>
<td>(r_A)</td>
<td>.88</td>
<td>.88</td>
<td>.88</td>
<td>.87</td>
<td>.80</td>
<td>.88</td>
<td>.85</td>
<td>.88</td>
<td>.94</td>
</tr>
<tr>
<td>(\mu_2)</td>
<td>.94</td>
<td>.94</td>
<td>.94</td>
<td>.93</td>
<td>.79</td>
<td>.94</td>
<td>.93</td>
<td>.94</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 19.
Product-Moment Correlation Coefficients Among Reliability Estimates for Random Data Sets (n = 17).

<table>
<thead>
<tr>
<th></th>
<th>KR20</th>
<th>$r_w$</th>
<th>$r_{H_H}$</th>
<th>$r_\gamma$</th>
<th>$r_H$</th>
<th>$r_R$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$r_A$</th>
<th>$\mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR20</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_w$</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{H_H}$</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_\gamma$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_t$</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_H$</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_R$</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.75</td>
<td>0.76</td>
<td>0.86</td>
<td>1.00</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.99</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Relationships between variables were also explored by comparing the relative rankings of computed reliability coefficients (Table 20). There was a general trend toward lower estimates of reliability by \( H_i \) and \( L_1 \). Highest coefficients were computed using \( L_2, \mu_2, \) and \( r_A \).

**Table 20.** Ascending Rank-Order of Mean Reliability Coefficients.

<table>
<thead>
<tr>
<th>Overall</th>
<th>Guttman Rows</th>
<th>Guttman Columns</th>
<th>Kuder-Richardson</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_t )</td>
<td>( H_t )</td>
<td>( H_t )</td>
<td>( H_t )</td>
<td>( H_t )</td>
</tr>
<tr>
<td>( L_t )</td>
<td>( L_t )</td>
<td>( L_t )</td>
<td>( L_t )</td>
<td>( r_{Hu} )</td>
</tr>
<tr>
<td>( r_H )</td>
<td>( r_W )</td>
<td>( r_W )</td>
<td>( r_{Hu} )</td>
<td>( L_t )</td>
</tr>
<tr>
<td>( r_W )</td>
<td>( r_{Hu} )</td>
<td>( r_{Hu} )</td>
<td>( r_{y} )</td>
<td>( r_W )</td>
</tr>
<tr>
<td>( KR20 )</td>
<td>( KR20 )</td>
<td>( KR20 )</td>
<td>( r_W )</td>
<td>( r_y )</td>
</tr>
<tr>
<td>( r_{y} )</td>
<td>( r_{y} )</td>
<td>( r_{y} )</td>
<td>( KR20 )</td>
<td>( KR20 )</td>
</tr>
<tr>
<td>( r_H )</td>
<td>( L_2 )</td>
<td>( r_H )</td>
<td>( r_H )</td>
<td>( r_H )</td>
</tr>
<tr>
<td>( r_R )</td>
<td>( \mu_2 )</td>
<td>( r_R )</td>
<td>( r_R )</td>
<td>( r_R )</td>
</tr>
<tr>
<td>( L_2 )</td>
<td>( r_H )</td>
<td>( L_2 )</td>
<td>( L_2 )</td>
<td>( L_2 )</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>( r_R )</td>
<td>( \mu_2 )</td>
<td>( \mu_2 )</td>
<td>( \mu_2 )</td>
</tr>
<tr>
<td>( r_A )</td>
<td>( r_A )</td>
<td>( r_A )</td>
<td>( r_A )</td>
<td>( r_A )</td>
</tr>
</tbody>
</table>

**Differences Between Estimation Techniques**

Since KR20 has been the most widespread technique for estimation of internal consistency reliability, differences between estimates by each method and KR20 were computed. The mean and standard deviation of differences are reported in Table 21. Reliability coefficients computed using the techniques of Loevinger, Guttman (\( L_t \)), Huck, and Winer, on average, returned estimates lower than KR20. The techniques
of Horst, Raju, Guttman ($L_z$), ten Berge and Zegers, and Ayabe generally resulted in estimates higher than KR20. Examination of the standard deviation of the differences between the respective methods and KR20 illustrates the consistency by which other estimates differed from KR20.

Table 21.

Differences in Reliability Coefficients

<table>
<thead>
<tr>
<th>Estimation Technique-KR20</th>
<th>$M_x$ (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>-.26 (.13)</td>
</tr>
<tr>
<td>$L_1$</td>
<td>-.07 (.03)</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>-.02 (.02)</td>
</tr>
<tr>
<td>$r_w$</td>
<td>-.02 (.01)</td>
</tr>
<tr>
<td>$r_H$</td>
<td>.00 (.02)</td>
</tr>
<tr>
<td>$r_{Ro}$</td>
<td>.04 (.04)</td>
</tr>
<tr>
<td>$r_R$</td>
<td>.12 (.04)</td>
</tr>
<tr>
<td>$L_z$</td>
<td>.14 (.11)</td>
</tr>
<tr>
<td>$\mu_z$</td>
<td>.15 (.11)</td>
</tr>
<tr>
<td>$r_A$</td>
<td>.26 (.16)</td>
</tr>
</tbody>
</table>

Excluded Data Matrices

Data were initially screened for the presence of negative average off-diagonal covariances in accordance with Kuder and Richardson's (1937) proposition that reliability was a function of positive inter-correlation. Negative average off-diagonal covariances were found in 24 data sets. Characteristics of the reliability estimates from the 24 matrices are summarized in Table 22. All reliability
Table 22.
Characteristics of Reliability Estimates from Data Matrices Excluded from Analysis.

<table>
<thead>
<tr>
<th></th>
<th>Maximum</th>
<th>Minimum</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR20</td>
<td>-0.06</td>
<td>-4.86</td>
<td>-0.36</td>
</tr>
<tr>
<td>$r_w$</td>
<td>-0.10</td>
<td>-5.11</td>
<td>-0.41</td>
</tr>
<tr>
<td>$r_{Hu}$</td>
<td>-0.10</td>
<td>-5.22</td>
<td>-0.40</td>
</tr>
<tr>
<td>$r_y$</td>
<td>-33.28</td>
<td>-186.74</td>
<td>-47.23</td>
</tr>
<tr>
<td>$H_i$</td>
<td>-0.01</td>
<td>-0.27</td>
<td>-0.05</td>
</tr>
<tr>
<td>$r_H$</td>
<td>-0.07</td>
<td>-5.67</td>
<td>-0.40</td>
</tr>
<tr>
<td>$r_R$</td>
<td>-0.10</td>
<td>-9.50</td>
<td>-0.62</td>
</tr>
<tr>
<td>$L_1$</td>
<td>-0.05</td>
<td>-4.25</td>
<td>-0.31</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.60</td>
<td>-2.35</td>
<td>0.20</td>
</tr>
<tr>
<td>$r_A$</td>
<td>0.79</td>
<td>-1.25</td>
<td>0.46</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>0.62</td>
<td>-2.28</td>
<td>0.22</td>
</tr>
</tbody>
</table>
estimates were negative in 4 (16.7%) of the 24 matrices. In
the remaining 20 matrices with negative off-diagonal inter-
correlations, all estimates were positive using Ayabe's $r_A$,
16 were positive with ten Berge and Zegers' $\mu_2$, and 15 were
positive as computed using Guttman's $L_2$.

The highest positive reliability estimate in a matrix
with an overall off-diagonal negative covariance was .79
computed using $r_A$. The lower limit of reliability was less
than -1.00, the theoretical minimum value for coefficients
of correlation, for all techniques except Loevinger's $H_k$.

Summary

The findings of this investigation may be summarized in
a series of bivariate plots illustrating the behavior of the
reliability estimates across a range of off-diagonal
covariance to total variance ratios. The plots provide
information about the upper and lower limits of the
reliability estimation techniques, linearity across the
covariance range, and associations among techniques.

Overall Data Set

Overall performance of the reliability estimation
techniques is illustrated in Figure 6. The contribution of
off-diagonal covariance to total variance in the data set
encompassed almost the entire spectrum from no covariance to
complete contribution of covariance.

Guttman's $L_1$, as the ratio of off-diagonal covariance
to total variance, served as a reference line within the
Figure 6.
Reliability Estimates as a Function of
Off-Diagonal Covariance: Total Sample.
figure. As an accepted conservative estimate of the lower limit, it was anticipated that "improved" reliability estimators would yield results higher than computed using \( L_1 \). With the exception of \( H_r \), most estimates exceeded \( L_1 \).

The theoretical lower limit of reliability was approached by all estimation techniques except for \( L_2, \mu_2, \) and \( r_A \). All techniques approach the theoretical upper limit of reliability except for \( L_1 \).

Kuder and Richardson’s formula 20, \( r_Y, r_w, \) and \( r_{Hu} \) were closely associated with each other and progressed smoothly across the range of covariance. There was more variability in the \( H_r, H, r_r, L_2, \mu_2, \) and \( r_A \) estimates.

Loevinger’s \( H_r \), Horst’s \( r_H \) and Raju’s \( r_R \) also progressed from near the theoretical minimum to the theoretical maximum. Estimates computed using the techniques of Ayabe (\( r_A \)), ten Berge and Zegers (\( \mu_2 \)), and Guttman (\( L_2 \)) failed to approach the theoretical minimum. Loevinger’s technique achieved both the upper and lower limits; however, the progression through the range of covariance was non-linear and characterized by wide variability in estimates.

**Guttman Data Matrices with Row Substitution**

Row substitution resulted in a small variation in the covariance proportion of total variance. Figure 7 illustrates data matrices confined to the upper middle portion of the overall range of the data set. No estimation technique approached the lower limit of reliability. Only
Figure 7.
Reliability Estimates as a Function of
Off-Diagonal Covariance: Guttman Sample with Row Substitution.
four techniques, $H_1$, $r_\gamma$, $r_H$, and $r_R$, achieved the upper limit. Overall, there was linear progression of reliability estimates throughout the covariance range. The slope of the $H_1$ estimates was visibly steeper. Excluding Loewinger's technique, estimates computed from the remaining formulas yielded reliability estimates .10 to .20 greater than $L_1$ and the slopes were generally parallel.

Guttman Data Matrices with Column Substitution

The effect of column substitution in a Guttman scale is shown in Figure 8. This cluster contained data matrices from across nearly the entire range of off-diagonal covariance ratios. The upper limit of reliability was attained by four techniques: $H_1$, $r_\gamma$, $r_H$, and $r_R$. The $L_2$, $\mu_2$, and $r_A$ techniques were characterized by failure to approach the anticipated minimum and more variability within the data cluster. Progression of other estimators through the range of covariances was essentially linear except for the $H_1$ method.

Kuder-Richardson Data Matrices with Column Substitution

The Kuder-Richardson cluster (Figure 9) was characterized by generally smooth progression across essentially the entire range of off-diagonal covariance:total variance ratios. The lower limit of reliability was approached by all estimating formulas except $L_2$, $\mu_2$, and $r_A$. Only $L_1$ did not approach the upper limit of reliability. The relationships among KR20, $r_w$, $r_Hu$, $r_H$, and
Figure 8.
Reliability Estimates as a Function of
Off-Diagonal Covariance: Guttman Sample with Column Substitution.
Figure 9.
Reliability Estimates as a Function of
Off-Diagonal Covariance: Kuder-Richardson Sample with Substitution.
$r_R$ remained unchanged. Estimates computed using the $H_i$ method were lower and the progression was non-linear.

Horst (1953) predicted that the difference between his method and KR20 would increase as variation in item difficulties increased. In the Kuder-Richardson cluster where there was no variation in item difficulty, the estimates obtained using Horst's method and KR20 were identical.

Random Data Matrices

Random matrices (Figure 10) were restricted to the lower two-thirds of the covariance:total variance spectrum. With most estimates there was a smooth progression through the range. The $L_2$, $\mu_2$, and $r_A$ formulas demonstrated the widest variability, did not approach the lower limit of reliability, and resulted in the highest reliability estimates. Progression of other estimates was smooth. The slope of the $H_i$ estimates was more gradual than the other estimates.

Of the 11 reliability estimates, only $r_r$, $r_R$, and $r_R$ met the criteria established for linearity, upper, and lower limits. Figure 11 illustrates the performance the these three techniques along with the differences in their computed values from KR20. The techniques of Raju and Horst provided estimates that were higher than KR20. Reliability coefficients computed with $r_r$ were comparable to those calculated with KR20.
Figure 10.
Reliability Estimates as a Function of Off-Diagonal Covariance: Random Substitution.
Figure 11.
Techniques for the estimation of internal consistency reliability have been debated for the past 50 years. The education and psychology literature contains numerous articles introducing new estimation methods or conducting limited comparisons of existing methods. There has not been a comprehensive systematic review of techniques for the estimation of internal consistency for unidimensional tests (Kuder, 1991).

Recent trends in psychometrics have focused on item response theory and sophisticated computer models for the estimation of reliability. These endeavors are important, but the role of traditional assessment of internal consistency should not be overlooked. The majority of tests are developed and administered for small-scale use. The most common example is a teacher-designed test for use in individual classrooms. Teachers, serving as test designers, need methods for the estimation of internal consistency and are served best by procedures that are easy to calculate with a minimum of computing resources. Traditional methods for the estimation of internal consistency meet these requirements. The overall purposes of this investigation were to identify the techniques available for internal consistency estimation and to compare selected techniques over a broad range of item-by-person matrices.
Reiteration of Findings

Techniques for the estimation of internal consistency reliability were to be judged as adequate if they met three criteria as proposed by Ayabe (1994), White and Saltz (1957), and the investigator: the upper limit of reliability must be unity, the lower limit of reliability must be zero, and progression between the two limits must be monotonic.

A fourth criteria, that there should be a strong association among techniques, was added for comparison. If all of these techniques are intending to measure the same concept, then they should be related to each other.

The Upper Limit of Reliability

Reliability estimates were computed for two data matrices that were theoretically perfect. All estimation techniques except Guttman’s $L_1$ achieved the 1.0 upper limit for the Kuder-Richardson data matrix. Only Horst’s $r_H$, Loevinger’s $H$, Raju’s $r_R$, and Cliff’s $r_\gamma$ estimates achieved unity in the Guttman data matrix. Failure to achieve unity as an upper limit for reliability may contribute to an underestimation of actual test reliability or a type II decision error: concluding that a test was not reliable when, in fact, it was.

The Lower Limit of Reliability

Overestimation of a test’s internal consistency may lead to a type I decision error: concluding that a test was
reliable when, in fact, it was not. Falsely concluding a test is reliable is probably the more detrimental type of psychometric error. Dramatically positive lower limits were projected through linear regression with Ayabe’s $r_A$, Guttman’s $L_2$, and ten Berge and Zegers’ $\mu_2$. A slightly positive lower limit was projected using Raju’s $r_R$ method. The 95% confidence interval for the regression Y-intercept included the ideal 0.00 with the $r_H$, KR20, and $r_\gamma$. The lower limit of reliability was slightly negative for the Winer and Huck methods. The most negative of Y-intercepts was computed when using Loevinger’s $H_t$ as the independent variable. The 95% confidence limit for $H_t$ was $-0.20, -0.08$.

**Goodness of Fit**

An estimation technique should produce a monotonic increase in reliability as the proportion of covariance increases. Simple regression with residual analysis and ANOVA were used to assess linearity and goodness of fit for the proposed estimation techniques.

For the three estimation methods meeting the upper and lower limit criteria ($r_\gamma$, $r_H$, and $r_R$), the proportion of explained variance, $R^2$, exceeded .97 in all clusters for the Horst and Raju methods. The overall $R^2$ for the $r_\gamma$ was .99; however, the value in the Guttman cluster with row substitution decreased to .92. Examination of the scatterplot of predicted and residual values suggested non-linearity.
Review of the scatterplot for the Loevinger estimation method also suggested a non-linearity. Goodness of fit tests for the Loevinger method remained statistically significant ($R^2 = .73$ for the overall sample), but lower than the other estimation methods.

Linearity is not an absolute requirement of a monotonic relationship. Graphic representation of generally increasing estimates with increasing off-diagonal covariance suggested a monotonic trend.

**Association Among Estimation Techniques**

In all sample clusters and the overall sample, there were statistically significant ($p < .001$) product-moment correlations among all estimation methods. The lowest of the significant associations were between $r_A$, $H_t$, and the other estimation methods.

The extent to which reliability estimates differed from the KR20 estimate was explored by computing difference scores. Of the three estimation techniques meeting the criteria for limits and linearity, the greatest increase in estimated reliability was computed with use of the Raju technique. There was a moderate increase in reliability estimates computed using Horst's technique in the Guttman clusters and minimal increase from the use of $r_\gamma$.

**Practical Interpretation of the Findings**

Of the 11 tested techniques for the estimation of internal consistency reliability, the use of the $r_A$, $L_2$, or
yielded higher estimates of reliability in situations where there was a smaller ratio of off-diagonal covariance to total variance. The potential benefits of higher reliability estimates was offset by the failure off all three techniques to achieve the upper limit of unity with the perfect Guttman pattern.

Performance of Guttman's $L_1$ could not be determined since it served as the point of reference for many of the comparisons. Historically $L_1$ has been considered an underestimate of true reliability. In the higher ranges of off-diagonal covariance ratios, $L_1$ values were typically lower than those of other techniques except for $H_1$. Guttman's $L_1$ was the only estimate failing to reach the theoretical maximum value in either the Guttman or Kuder-Richardson model.

The techniques of Huck, Kuder-Richardson, and Winer produced reliability estimates similar to each other. Although the methods of Winer and Huck may be more accurate conceptualizations of psychometric models, there was essentially no practical advantage to their use. The Kuder-Richardson method is easy to compute manually or using widely available computer software. These three techniques were also characterized by a failure to achieve the upper limit of reliability in the perfect Guttman matrix.

Three techniques met the criteria for upper and lower limits. As proportional off-diagonal covariance increased,
there were linear increases in estimated reliability using the Cliff's \( r_\gamma \), Horst's \( r_H \), Raju's \( r_R \) techniques. Although \( r_\gamma \) met all the requirements, there was essentially no numerical advantage to using this estimator as opposed to the readily available KR20. Horst's \( r_H \) provided reliability estimates higher than KR20 when test items varied in difficulty. Only Raju's \( r_R \) consistently provided estimates higher than KR20.

The curvilinear performance of Loevinger's \( H_i \) presents a unique problem. The lack of linearity initially seems undesirable; however, this may not be the case. Across a wide range of off-diagonal covariance to total variance ratios, \( H_i \) estimates are low leading the test designer to conclude that the test may be unreliable. Proportionally higher covariances dramatically improve the \( H_i \) estimate creating an almost dichotomous situation for interpretation: the reliability statistic being either completely acceptable or unacceptable.

Relation of Findings to Research Design

Investigations concerning internal consistency reliability are handicapped because the dependent variable, true reliability, is latent. Reliability was operationally defined as the ratio of off-diagonal covariance to total variance. Selection of the off-diagonal covariance to total variance ratio as the independent variable for simple regression and bivariate comparisons contributed to the
linear performance of methods incorporating the same operational definition of reliability. The linear performance of $L_1$ and KR20 were direct results of the experimental design.

Relation of Findings to Review of Literature

The introduction of Kuder and Richardson's (1937) formula 20 was the starting point for the estimation of reliability using one complete test and its hypothetical equivalent. The widespread use of their technique also qualified KR20 as the prototype for comparison with other reliability estimation methods. Predominant criticisms of KR20 are the use of $\phi$ to approximate the correlation between an item with its hypothetical equivalent and the attenuated upper limit of reliability.

Limitations of Phi

Prior to Kuder and Richardson, test reliability was measured by computing the correlation between two separate tests. Instead of using two real tests, Kuder and Richardson (1937) devised a method for correlating one real and one hypothetical test. Kuder and Richardson proposed that the average off-diagonal covariance of the variance/covariance matrix for the test was an approximation of the intra-item covariance for the real test. Correlations between the comparable items on the real and hypothetical tests were represented on the diagonal of the variance-covariance matrix. The on-diagonal elements could
not be directly calculated since they relied upon the relationship with a hypothetical test. Kuder and Richardson proposed that if a test were unidimensional, each item measured the same concept as the total test and the correlation between the individual item and the total test was used to represent the like-item correlation. The result of this assumption was the substitution of the average inter-item correlate for the on-diagonal elements. Johnson (1945) and Loevinger (1947) argued that Kuder and Richardson’s substitution strategy led to an underestimate of reliability because $\phi$ could reach its maximum only when items were all of equal difficulty. If item difficulties were not equal, test reliability would be underestimated.

As an alternative to $\phi$, Cliff (1984) proposed the use of $\gamma$. According to Cliff, the use of $\gamma$ would improve reliability since $\gamma$ was not affected by item difficulty. In this investigation, reliability was minimally improved by the use of Cliff’s $r_{A'}$ ($M_{\text{diff}} = .001, S = .02$).

Ayabe (1994) proposed that the correlation between any item and its hypothetical counterpart should be at least as good as the correlation between the real item and any other real item. He suggested using the highest covariate in lieu of the average of covariates. There was a substantial increase in reliability estimates calculated using $r_{A'}$ ($M_{\text{diff}} = .26, S = .16$).
Horst (1953), Loevinger (1947), and Raju (1982) used adjusted ratios as reliability estimates. Recognizing the limitations of $\phi$, they developed estimates that compared calculated values to their theoretical maximums.

The use of Loevinger's method resulted in estimates of reliability lower than calculated by other methods. Cudeck (1980) suggested that $H_t$'s in the range of .3 to .5 corresponded to an acceptable range of reliabilities. It was difficult to replicate Cudeck's finding in this investigation because of fluctuation in $H_t$ associated with variation in the range of item difficulties.

Horst (1953) modified $H_t$ by adding the ratio of maximal to total test variance. Horst's correction resulted in less sensitivity to ranges of item difficulties and contributed to resolution of the non-linearity observed with the use of $H_t$.

Raju (1982) combined the concept of comparing actual to maximal values with Kuder and Richardson's estimation technique. His $r_R$ was the ratio of actual to maximal KR20 estimates. Reliability estimates using $r_R$ were greater than calculated by Horst or Kuder and Richardson ($M_{\text{diff}} = .12, S = .04$).

Cronbach and Azuma (1962) and Guilford and Fruchter (1973) commented that no rationale was presented by Loevinger (1947) or Horst (1953) for the use of actual to maximal ratios (Raju's technique was not published until
1982). Results from Horst’s and Raju’s ratio-based methods corresponded to values obtained from techniques with a direct basis in classical measurement theory.

**Limits of Reliability**

The attenuated upper limit of reliability estimates calculated using KR20 has been documented (Johnson, 1945; Loevinger, 1947). Upper limit attenuation is not limited to KR20. The techniques of Ayabe (1994), Guttman (1945), Huck (1978), ten Berge and Zegers (1978), and Winer (1971) also failed to reach the theoretical upper limit when computed for a perfect Guttman scalogram.

The lower limit of reliability has not been documented in previous investigations. Through linear regression, lower limits of reliability were projected in this study. Guttman’s $L_2$, $\mu_2$, and $r_A$ had positive $Y$-intercepts that were different from zero both statistically ($p = .001$) and practically. Of less practical significance were the deviations projected for $H_i$ and $r_X$.

Discussion of the lower limit of reliability is also obscured by the problem of negative reliability estimates when the average off-diagonal covariances is negative. Kuder and Richardson (1937) described the situation as "inadmissible" (p. 103). Twenty-four data matrices were excluded from this investigation on the basis of negative off-diagonal covariances. Ideally, an alternative to KR20 would not be vulnerable to the calculation of negative
reliability estimates. None of the 11 estimation techniques compared in this investigation computed positive estimates for all of the 122 original matrixes. Positive estimates of reliability were computed using Ayabe's $r_A$ in 118 of the 122 matrixes. The minimum value for reliability estimates was less than -1.00 for all estimation techniques except $H_i$.

Two interpretations may be proposed for dealing with negative reliabilities. If Kuder and Richardson's viewpoint is accepted, then the calculation of a negative value warns the test designer of the violation of an assumption cardinal to the use of the estimation technique.

Conversely, if it is argued that an estimation technique should be capable of providing an estimate under in any testing situation, then one of the potential advantages to a KR20-alternative would be its versatility. No estimation technique provided a positive reliability estimate in the complete set of 122 data matrices. The technique with the highest frequency of positive reliability estimates, $r_A$, failed to achieve 1.0 as a maximum value.

**General Criticisms**

Huck (1978) and Winer (1971) attempted to increase reliability estimates by further partitioning of the error associated with test scores as described in classical measurement model. Huck's proposal advanced the conceptualization of the mathematical model of test scores;
however, the practical improvement in estimates of reliability was negligible.

In 1945, Guttman proposed a series of lower bounds for reliability. His $L_1$, which was the proportion of total variance not attributable to items served as a reference point for this investigation. Adding the sum of squares of the covariances to $L_1$ ($L_2$), was reported to be a better estimate of reliability. ten Berge and Zegers (1978) demonstrated that reliability estimates can be increased further by continuously increasing the power of the covariances without danger of exceeding the true reliability. Estimates of reliability were increased in their investigation by using this technique; however neither $L_2$ or ten Berge and Zegers' $\mu_2$ approached the lower limit of reliability.

**Standards for Comparison of Estimation Methods**

Proponents and critics have explored the characteristics of various estimation methods, yet understanding the nuances of each of the techniques is irrelevant unless certain performance criteria are met. Ayabe (1994) and White and Saltz (1957) provided standards for any reliability estimation technique. Of the 11 estimation methods testing in this investigation, only $r_r$, $r_H$, $r_R$ met the criteria of linear progression between an upper limit of unity and a lower limit of zero.
Conclusions

This investigation evaluated techniques for the estimation of reliability in unidimensional tests with dichotomously-scored items. Only methods that were considered to be easily calculated by individuals with minimal resources were included. Based on the findings of this investigation, it was concluded that:

1. The majority of estimation procedures do not meet minimal criteria for interpretation. Without clearly defined upper and lower limits, merits of any particular technique are moot.

2. Only Cliff’s $r_\gamma$, Horst’s $r_H$, and Raju’s $r_R$ met the criteria of linear progression between an upper limit of unity and a lower limit of zero, based on the criteria used for selection in this investigation.

3. Substitution of $\phi$ with $\gamma$, an estimate of association unaffected by variable distribution produced nominal change in reliability estimates.

4. Ratio-based estimates of reliability were statistically related to conceptually-based estimates.

Significance of the Investigation

The goal of this investigation was to identify internal consistency reliability estimation techniques that provided reasonable results with minimal computational effort. Three reliability estimation techniques were proposed to replace KR20. Gamma reliability ($r_\gamma$) produced estimates that were
similar to KR20, except that unity was achieved as an upper limit. The foundation for $r_\gamma$ is reasonable. Cliff (1984) substituted $\gamma$ for $\phi$ in Kuder and Richardson's conceptual formula. The use of Horst's $r_H$ resulted in estimates of reliability averaging .04 greater than that computed using KR20. Although the conceptual foundation of techniques incorporating ratios of actual to maximal variances have been challenged by Cronbach and Azuma (1962) and Guilford and Fruchter (1973), this investigation demonstrated close statistical association between ratio-based techniques and conceptually-based methods. Raju's $r_R$ produced reliability estimates averaging .12 greater than KR20. In addition to criticism of the method as ratio-based, the 95% confidence interval of the estimates lower limit was slightly greater than zero (.07, .10).

Kuder and Richardson's formula 20 should be used with caution. Attenuation of the upper limit of reliability when there is a dispersion of item difficulties may lead a test designer to underestimate internal consistency reliability. Only minimal extra calculation is required to increase the estimated reliability by computing $r_H$ or $r_R$. Conversely, the widespread availability of KR20 makes it an attractive option. The performance of KR20 is such that interpretation of estimates is reasonably evident. If a negative or low positive value is computed, the test developer should re-evaluate the test. If the computed estimate is judged as
satisfactory, then it reasonable to conclude that the test is internally consistent. The most challenging circumstance is the situation where the computed estimate is marginally acceptable. Especially if the range of item difficulties is wide, the test designer may consider using another technique, such as Cliff’s $r_\gamma$, Horst’s $r_H$, or Raju’s $r_R$.

The most surprising finding of the investigation was the performance of most of the estimation techniques at lower levels of reliability. It can be argued that performance of estimators in low reliability situations is not important because the computation of a low value would immediately warn the test constructor of testing problems, but there are two reasons why this may not be true. First, the slope of a line with a maximum value of 1.0 and a positive Y-intercept is more gradual than a line with a 0.0 Y-intercept. At lower levels of reliability that magnitude of the difference between the two lines is greater than at higher levels of reliability contributing to a progressive increase in overestimation as actual reliability decreases. Second, the conservative nature of computed estimates suggests that true reliability is somewhat greater than the computed value. The higher estimate resulting from the slope of the regression line may provide a false sense of security to the test constructor who assumes that true reliability is at least as great as the computed value.
Implications for Future Research

Most importantly, this investigation has suggested that reliability estimation using classical methods is not an issue that has been resolved. Kuder (1991) was correct in encouraging continued exploration of the topic. In the process of comparing techniques for the estimation of internal consistency, a critical weakness in the literature has been discovered: lower limits of reliability. Investigations in the future should document the complete range of reliability.

This investigation maintained a narrow descriptive focus: general performance of easily-computed estimation techniques. Methods such as the algorithm-based MSPLIT, computer search models, and techniques for reliability estimation in multi-dimensional tests were not included. Continued application of the methodology used in this investigation in these additional circumstances could provide additional insight into the estimation of reliability.

The set of data matrices used for this study provided a convenient method for manipulation of a test's variance/covariance matrix. Future investigations could use this system for a more complete analysis to include variations in the shape of the distribution, overall test difficulty, and numbers of test items.
References


La Duca, A. (Ed.). (1994). Defining an agenda for validation research for professional licensure and certification examinations [Special issue]. *Evaluation and the Health Professions, 17(2)*.


