Coefficient Alpha and Related Internal Consistency Reliability Coefficients

H. G. Osburn
University of Houston

The author studied the conditions under which coefficient alpha and 10 related internal consistency reliability coefficients underestimate the reliability of a measure. Simulated data showed that alpha, though reasonably robust when computed on \( n \) components in moderately heterogeneous data, can under certain conditions seriously underestimate the reliability of a measure. Consequently, alpha, when used in corrections for attenuation, can result in nontrivial overestimation of the corrected correlation. Most of the coefficients studied, including \( \lambda_2 \), did not improve the estimate to any great extent when the data were heterogeneous. The exceptions were stratified alpha and maximal reliability, which performed well when the components were grouped into two subsets, each measuring a different factor, and maximized \( \lambda_4 \), which provided the most consistently accurate estimate of the reliability in all simulations studied.

Coefficient alpha is an appropriate reliability estimator for composite measures containing multiple components. A component may be a test item, a judge, a Thematic Apperception Test (TAT) card, a survey question, a subtest, or a test that is being combined into a composite test battery. Multiple components may be homogeneous in the sense of measuring a single latent variable, or they may be heterogeneous in the sense of measuring two or more factors or latent variables. Because of coefficient alpha's flexibility, its use is ubiquitous in most areas of psychology as well as in many other disciplines. For example, Cortina (1993) reported that a review of the Social Sciences Citation Index for the period from 1966 to 1990 revealed that Cronbach's 1951 article on coefficient alpha had been referenced approximately 60 times per year in a total of 278 different journals.

Although equivalent-forms estimators are often preferred, they are appropriate only in very specialized circumstances in which two or more test forms constructed to be parallel are available. Test-retest coefficients are problematical for paper-and-pencil tests because of memory and practice effects. Internal consistency reliability estimators are very flexible and appropriate for a wide variety of circumstances in which an estimate of reliability is needed. Examples are (a) a single administration of a composite test containing \( n \) components, (b) two or more subtests combined to form a composite score, (c) two or more tests combined in a battery score, and (d) two or more judges' ratings combined into a composite.

Several investigators have noted that internal consistency reliability coefficients, including coefficient alpha, based on a single administration of a test may overestimate reliability because such coefficients assign transient error due to differences in test administration, temporary changes in the examinee, and so on to true-score variance (Guion, 1965; Schmidt and Hunter, 1996). In addition, if the errors made in responding to test items are positively correlated, coefficient alpha may be inflated (Komaroff, 1997). Also, reliability may be overestimated if coefficient alpha is erroneously used to estimate the reliability of a speeded test. Although it is true that coefficient alpha may sometimes be inflated because of the conditions mentioned above, a potentially more serious problem is the tendency of coefficient alpha to underestimate the true reliability when the data are multidimensional. If the items of a test conform to classical test-theory assumptions, coefficient alpha is a lower bound to the true reliability of a measure and frequently is an underestimate of reliability if the measure contains a small number of heterogeneous items. This point was made as far back as 1945 by Guttman,
Muchinsky (1996) argued that correcting for errors of measurement in the criterion remains controversial in meta-analyses, criterion-related validity studies, and utility analyses. Finally, Schmidt and Hunter (1996), in a discussion of using various reliability coefficients in corrections for attenuation, largely ignored the tendency of coefficient alpha to underestimate the reliability because of item heterogeneity. A second purpose of this study was to examine the extent to which correlation coefficients, corrected for unreliability, are overestimated when using coefficient alpha to estimate reliability.

Internal Consistency Reliability Estimators

Although coefficient alpha is widely used in practical work, a significant body of literature has developed in which several less familiar internal consistency reliability coefficients have been proposed. All of these coefficients have in common an attempt to reduce the underestimation of reliability when the components of a measure are not equivalent. All except standardized alpha and maximal reliability are discussed by Feldt and Brennan (1989). However, in the interests of clarity and reader convenience, a brief discussion of each estimator's important properties is presented in this section, along with a computing formula for each coefficient.

It should be noted that this article does not specifically address the generalizability approach to reliability, which goes far beyond classical reliability theory. However, in certain designs, such as a $p \times i$ single-facet study, the generalizability coefficient is the same as coefficient alpha, so the findings of the present study apply.

Coefficient Alpha

In the theory of composite tests, the reliability of a measure, $\rho^2_{XT}$, is defined as the ratio of the variance of the true scores ($T$) of the measure to the variance of the observed scores ($X$). In the framework of classical reliability theory and for clarity, we call $\rho^2_{XT}$ the true reliability of a composite measure. As mentioned above, coefficient alpha is a lower bound to the true reliability (Guttman, 1945). However, under very restricted conditions coefficient alpha is equal to the true reliability. In his derivation of the formula for coefficient alpha, Gulliksen (1950) considered two $n$-item tests that are parallel item-for-item and assumed that the average covariance of an item in one test with the parallel item in the other test is equal to the av-

Corrections for Attenuation

If reliability is underestimated, attenuation is overestimated (Guttman, 1945). Because of their tendency to underestimate the reliability of heterogeneous tests, internal consistency estimators have been especially controversial when used in correction for attenuation formulas (Cureton, 1958; Lord & Novick, 1968; Nunnally, 1978). The use of such reliability estimates to correct for attenuation will in many cases result in overestimation of the correlation between the true scores on the two measures and in some cases may result in substantial overestimation. In a recent article, Muchinsky (1996) argued that correcting for errors of...
Average covariance of pairs of items within a single test, that is,

$$\Sigma \rho_{ij} \sigma_i^2 = \Sigma \rho_{ij} \sigma_i \sigma_j / (n - 1), \quad i \neq j. \quad (1)$$

where $\rho_{ij} \sigma_i^2$ is the covariance between parallel items and $\rho_{ij} \sigma_i \sigma_j$ is the covariance between nonparallel items.

If Equation 1 holds, then coefficient alpha is equal to the true reliability. Novick and Lewis (1967) showed that Equation 1 holds if and only if the components of the test are essentially tau equivalent, that is,

$$T_i = a_i + T_j, \quad (2)$$

for all $i$ and $j$.

Tau-equivalent measures have equal true scores, and as indicated in Equation 2, essentially tau-equivalent measures have true scores that differ only by an additive constant. This article both essential tau equivalence and tau equivalence lead to the same conclusions. Therefore, essential tau equivalence will be referred to simply as tau equivalence. Both tau equivalence and essential tau equivalence imply that the correlations between true scores are all unity and that the variances and covariances of the true scores on the components of the measure are all equal. Tau equivalence also implies that the observed component intercorrelations corrected for attenuation are unity and that the measure is unidimensional with equal factor loadings but possibly unequal error variances.

There are two important advantages of a test whose components are tau equivalent: (a) The test can be interpreted as measuring a single underlying construct and (b) coefficient alpha computed on the components of the test is exactly equal to the true reliability.

Congeneric equivalence is closely related to tau equivalence. If the components of a measure are congeneric equivalent, the correlations between the true scores are unity but variances of the true scores may vary (Jöreskog & Sörbom, 1990). Congeneric equivalence implies that the measure is unidimensional, but the factor loadings and error variances may vary. Congeneric tests share the two advantages of tau-equivalent tests because such tests measure a single underlying construct, and the data presented in this article show that coefficient alpha for such tests is reasonably close to the true reliability of the composite measure.

Finally, if the components of a measure are parallel, the correlations between true scores are all unity and the variance and covariance of the component true scores of the measure are all equal. In addition, the components are equally reliable. Parallel components are unidimensional with equal factor loadings and equal error variances.

Coefficient alpha computed on heterogeneous components may seriously underestimate the true reliability. This issue is closely related to the multidimensionality of the latent variables underlying the components. If the components are unidimensional, coefficient alpha will be an adequate estimate of the true reliability in all cases. If the components are characterized by two or more moderately correlated factors (i.e., in the range of .2-.4), coefficient alpha may be seriously underestimated.

Coefficient alpha may be computed using variance components but is ordinarily computed by the following equation:

$$\alpha = n(1 - \Sigma \sigma_i^2 / \sigma_T^2) / (n - 1), \quad (3)$$

where $\sigma_i^2$ is the variance of the $i$th component, $\sigma_T^2$ is the variance of the test, and the total score on the test is the sum of the $n$ component scores. It is important to note that coefficient alpha can be computed on $n$ components of a measure, the $n$ components grouped into split halves, or the $n$ components grouped into three or more parts. In all of these applications coefficient alpha is still a lower bound to the true reliability (Guttman, 1945).

An alternate formula for coefficient alpha is

$$\alpha = n^2 \text{Ave}(\sigma_i^2) / (n \text{Ave}(\sigma_i^2) + n(n - 1) \text{Ave}(\sigma_i^2)), \quad (4)$$

where $\text{Ave}(\sigma_i^2)$ is the average covariance among components and $\text{Ave}(\sigma_i^2)$ is the average variance of the components in the measure. Several conclusions about coefficient alpha can be drawn from this formula: (a) If $\text{Ave}(\sigma_i^2)$ is large relative to $\text{Ave}(\sigma_i^2)$, coefficient alpha will be large; (b) if there are many components in the measure, the $n(n - 1)$ average covariances will dominate the $n$ average variances and coefficient alpha will tend to be large; (c) if $\text{Ave}(\sigma_i^2)$ is small relative to $\text{Ave}(\sigma_i^2)$ either because the covariances are uniformly low or because the covariances are very variable, coefficient alpha will tend to be low to moderate.

Raju's Coefficient

Raju (1977) derived a formula for estimating reliability when different numbers of items are assigned to parts of a test so that the parts are at most congeneric. Raju's coefficient is defined by
\[ p = \frac{(\sigma_1^2 - \Sigma \sigma_j^2) \cdot ((1 - \Sigma \lambda_i^2) \sigma_x^2)}{\Sigma \lambda_i^2}, \]  
(5)

where \( \lambda_i = n_i/\Sigma n_i \) and \( n_i \) is the number of items in the \( i \)th part. If the parts contain equal numbers of items, all \( \lambda_i \) equal 1/k (where k is the number of parts) and Raju’s coefficient equals coefficient alpha.

**Angoff–Feldt Coefficient**

Feldt (1975) derived this coefficient for the situation in which a measuring instrument could be divided into only two parts of arbitrary length. He assumed that the available parts were homogeneous in content but that because of the unequal length of the parts the scores could only be congeneric equivalent. He also made the additional assumption that the two parts were differentially shortened versions of the total test. This latter assumption amounts to assuming that the sum of the error variances for the two parts is equal to the error variance of the total test. Feldt’s coefficient will exceed coefficient alpha whenever the part-test variances are unequal. The coefficient is not a lower bound to the true reliability and in some circumstances may overestimate the true reliability. Angoff (1953) and Kristof (1971) have also derived versions that ultimately lead to the same formula. Feldt and Brennan (1989) have designated this coefficient as the Angoff–Feldt coefficient:

\[ p = 4\sigma_{ab}/(\sigma_x^2 - (\sigma_a^2 - \sigma_b^2)/\sigma_x^2). \]  
(6)

The Angoff–Feldt coefficient is a special case for two-part splits of the Feldt coefficient for multipart splits presented below.

**Feldt’s Coefficient**

Feldt and Brennan (1989) included the following coefficient, due to Feldt:

\[ p = \frac{(\sigma_x^2 - (\sigma_a^2 - \Sigma \sigma_j^2))}{(\sigma_x^2 - \Sigma \sigma_{ij}^2)}. \]  
(7)

Feldt’s coefficient is based on the premise that \( \lambda_i \) in Raju’s coefficient could be estimated by \( \Sigma \sigma_i/\sigma_x^2 \). The Angoff–Feldt coefficient is a special case of the Feldt coefficient when components are divided into only two parts.

**Kristof’s Coefficient**

Kristof (1974) derived a reliability estimator for a test divided into three parts, assuming the parts are congeneric equivalent:

\[ p = \frac{(\sigma_{12} \sigma_{13} + \sigma_{12} \sigma_{23} + \sigma_{13} \sigma_{23})}{\sigma_{12} + \sigma_{13} + \sigma_{23}} \]  
(8)

where \( \sigma_{12}, \sigma_{13}, \) and \( \sigma_{23} \) are the covariances among the three parts and \( \sigma_x^2 \) is the total test variance. If the three parts are congeneric, the coefficient will give the precise value of the reliability coefficient even though the true-score variances of the parts differ. However, if the three splits are not congeneric equivalent, underestimation will occur. Kristof pointed out that the coefficient will always be at least as accurate as coefficient alpha computed on the three parts. Kristof’s coefficient is a special case for three-part splits of the Feldt–Gilmer coefficient for multipart splits presented below.

**Feldt–Gilmer Coefficient**

Gilmer and Feldt (1983) have proposed the following coefficient for multipart tests:

\[ \text{Feldt–Gilmer} = \frac{((\Sigma D_i)^2/((\Sigma D_i)^2 - \Sigma D_i^2))(\sigma_x^2 - \Sigma \sigma_j^2)}{\Sigma \sigma_{ij}^2}; \]  
(9)

\[ D_i = (\Sigma \sigma_{ij} - \sigma_i - \sigma_j^2/(\text{sum is over j}). \]  
(10)

where Row i is the row with the largest sum of interpart covariances. When \( i = l \), \( D_l = 1.0 \). Feldt–Gilmer is a generalization of Kristof’s coefficient for tests containing three or more parts. Thus, the Feldt–Gilmer and Kristof’s coefficients give equivalent results for three-part splits. This coefficient assumes congeneric equivalence.

**Lambda2**

Lambda2 (Guttman, 1945) is also a lower bound to the true reliability and equals the true reliability if the components are tau equivalent. Lambda2 is interesting because it always gives a lower bound that is as good as coefficient alpha but in some circumstances may be considerably better. Lambda2 is computed by the following equation:

\[ \lambda_2 = 1 - \Sigma \sigma_j^2/\sigma_x^2 + \sqrt{(n \Sigma \Sigma \sigma_{ij}^2/(n-1))/\sigma_x^2}, \]  
(11)

where \( \sigma_{ij} \) is the covariance between items i and j.

Ten Berge and Zegers (1978) have developed an infinite series of lower bounds to the true reliability: \( u_0, u_1, u_2, u_3 \), and so on; \( u_0 \) is coefficient alpha, \( u_1 \) is Lambda2, and

\[ u_2 = (\Sigma \Sigma \sigma_{ij} + \sqrt{\Sigma \Sigma \sigma_{ij}^2 + \sqrt{n/(n-1)\Sigma \Sigma \sigma_{ij}^2}})/\sigma_x^2, \]  
(12)
Ten Berge and Zegers computed $u_0$, $u_1$, and $u_2$ on college test data. In their data for a four-item test (similar to our four-component simulations), coefficient alpha was .4359, $\lambda_2$ was .4602, and $u_2$ was .4673. It was decided not to include data on $u_2$ in this article because of the small difference between $u_2$ and $\lambda_2$.

**Maximized Lambda4**

Cronbach (1951) showed that coefficient alpha is the expectation of coefficient alpha computed on all possible $1/2(2n)!/(n!)^2$ split halves of a test containing $2n$ items. If the data are tau equivalent, all coefficient alphas computed on the split halves will be equal to coefficient alpha computed on the $2n$ items. However, if the data are not tau equivalent, some split halves will give higher reliability estimates than coefficient alpha computed on the $2n$ components of the measure. Because coefficient alpha is a lower bound to the true reliability, coefficient alpha computed on some split halves may provide lower bounds that are closer to the true reliability. This property was pointed out by Guttman (1945) and used by Callender and Osburn (1977) in an attempt to improve reliability estimates for corrections for attenuation. The term $\lambda_4$ was introduced by Guttman. Lambda4 is the same as coefficient alpha computed on split halves of the test:

$$\lambda_4 = 2(1 - (\sigma_a^2 + \sigma_b^2)/\sigma_x^2),$$

(13)

where $\sigma_a^2$ and $\sigma_b^2$ are the variances of each half and $\sigma_x^2$ is the total variance.

It is not always feasible to compute lambda4 on all possible splits. For example, a 50-item test can be divided into two 25-item half tests in over 63 trillion ways. However, for a small number of items lambda4 can be computed for all possible splits to find maximized lambda4, and for a moderate number of components an approximation to the largest lambda4 can often be obtained. In this study, for measures with four components lambda4 was computed for all three possible splits, and for measures with eight components an approximation to maximized lambda4 was computed for all possible splits. Feldt and Brennan (1989) gave the following formula for stratified alpha:

$$\text{Stratified } \alpha = 1 - \sum \sigma_i^2(1 - \alpha_i)/\sigma_x^2,$$

(14)

where $\alpha_i$ is coefficient alpha computed on the $i$th subtest and $\sigma_i^2$ is the variance of the $i$th subtest. It is clear from inspection of the above formula that if the correlations among components in the same subtests are high and the correlations among components in different subtests are low, stratified alpha will be a better estimate than coefficient alpha.
Standardized Alpha

Use of standardized alpha implies that observed scores on the components are standardized before summing. However, it is important to note that the true reliability of a measure is the same for both standardized and unstandardized observed scores. Standardized alpha is not a lower bound to the true reliability and may yield estimates that are either lower or higher than the true reliability. When the components of a composite measure are congeneric, standardized alpha will always exceed the true reliability. Also, if the components are tau equivalent and reliabilities of the components differ, standardized alpha will always exceed the true reliability. Standardized alpha is computed using the correlations rather than the covariances:

\[ \text{Standardized } \alpha = n(1 - \frac{n}{n + \sum \Sigma \rho_{ij}}/(n - 1)), \quad i \neq j, \]  

where \( \rho_{ij} \) is the correlation between the \( i \)th and \( j \)th items. It is easily shown that standardized alpha is the same as the generalized Spearman–Brown formula for lengthening a test \( n \) times when the average correlation (\( \overline{p} \)) is used in the formula, that is,

\[ \text{Standardized } \alpha = n\overline{p}/(1 + (n - 1)\overline{p}), \]  

where \( \overline{p} \) is the average of the \( n(n - 1) \) correlations.

Maximal Reliability

Maximal reliability provides an estimate of reliability when the components in a test can be grouped into subsets. Maximal reliability (\( R_K \)) assumes that (a) all components within a subset are parallel, that is, the components in a subtest have equal reliabilities and equal variances; and (b) the components in different subtests may have differing reliabilities and variances. If these two assumptions are met, the components in all subtests are congeneric (Li, Rosenthal, & Rubin, 1996). Maximal reliability is a generalization of the Spearman–Brown approach and provides a component-weighting procedure to achieve maximal reliability. Maximal reliability is easy to compute using a hand calculator:

\[ R_K = \frac{A}{K(1 + (K - 1)p) + A}, \]  

where \( K \) = number of subsets or subtests, \( A = n_1\rho_1/((1 - \rho_1) + n_2\rho_2/(1 - \rho_2) + \ldots + n_k\rho_k/(1 - \rho_k)), \) \( \rho \) is the common correlation among the subsets, \( n_i \) is the number of type \( i \) components, and \( \rho_i \) is the reliability of a type \( i \) item. For two subtests, \( \rho \) can be estimated by

\[ r_{12}/(r_{11}r_{22}), \]  

where \( r_{12} \) is the average correlation between Type 1 components and Type 2 components. For more than two subtests, \( \rho \) can be estimated for each pair of subsets and then averaged.

Study of Coefficient Alpha and Related Reliability Estimates

This study focused on two types of demonstration data. The first data type simulated measures with four components and relatively large correlations among the components (\( \rho \approx 0.50 \)). Examples of this data type include supervisor ratings, cognitive ability subtests, and personality dimensions. The second data type simulated measures with eight components and relatively low correlations among the components (\( \rho \approx 0.25 \)). Examples of this data type include questionnaire measures of organizational commitment, role ambiguity, personal preferences, perception, and personality facets. Various scenarios were simulated, and the extent to which the various reliability estimators underestimated or overestimated the true reliability was determined.

Demonstration Data

Demonstration data were created to study estimation of true reliability by the various internal consistency coefficients. The general method was to generate simulated population covariance matrices for both true and observed component scores. From these data the true reliability and 11 estimators of reliability could be computed for varying degrees of heterogeneity among the components. A description of the simulation method is in the Appendix.

The following data sets were simulated: (a) homogeneous, one-factor data with equal factor loadings and equal error variances (parallel components); (b) homogeneous, one-factor data with equal factor loadings but unequal error variances (tau-equivalent components); (c) homogeneous, one-factor data with unequal factor loadings and unequal error variances (congeneric-equivalent components); and (d) heterogeneous, two-factor data with decreasing correlations between the factors (heterogeneous components).

Three types of situations with which an investigator might be confronted were simulated: (a) ungrouped data such as a test containing \( n \) items, (b) data grouped into two subsets such as two subtests of a larger test, and (c) data grouped into three subsets. If the data are ungrouped, an investigator has a choice of computing a wide variety of reliability coefficients by...
analyzing the ungrouped data as is or by grouping the data in various ways to find the largest lower-bound reliability estimate. As shown in this demonstration, an improved lower-bound estimate can sometimes be obtained by grouping the data and computing a reliability estimate on the grouped data. If the data are grouped to start with, an investigator still has a number of lower-bound reliability estimates to choose from. In the case of two groups, six different estimates are available. The downside in computing a large number of the lower-bound alternatives to coefficient alpha is the possibility of capitalizing on sampling error. However, as shown in this demonstration, many of the alternative coefficients give similar results, so it is not necessary to compute a large number of estimates.

Results for Four-Component Data

Analysis of four components ungrouped. As can be seen from Table 1, when the components were parallel, all coefficients were equal to the true reliability. When the components were tau equivalent, all coefficients were equal to the true reliability except for standardized alpha, which overestimated the true reliability by a small amount. When the components were congeneric, only Feldt-Gilmer and maximized lambda4 were equal to the true reliability, although coefficient alpha and lambda2 were very close, and standardized alpha and the Feldt coefficient slightly overestimated the true reliability. When the components were generated by two factors, all coefficients except maximized lambda4 underestimated the true reliability to a greater or lesser extent. For example, when the components were moderately heterogeneous coefficient alpha was about 92% of the true reliability, and when the components were strongly heterogeneous coefficient alpha was about 78% of the true reliability. In the strongly heterogeneous case the correlation between the two factors was only .2. This degree of heterogeneity is probably rare in practical applications.

<table>
<thead>
<tr>
<th>Reliability coefficient</th>
<th>Parallel</th>
<th>Tau eq.</th>
<th>Congen.</th>
<th>Slight</th>
<th>Moderate</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>True reliability</td>
<td>.800</td>
<td>.798</td>
<td>.786</td>
<td>.781</td>
<td>.760</td>
<td>.703</td>
</tr>
<tr>
<td>Four components ungrouped</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.800</td>
<td>.798</td>
<td>.778</td>
<td>.752</td>
<td>.696</td>
<td>.547</td>
</tr>
<tr>
<td>Feldt</td>
<td>.800</td>
<td>.798</td>
<td>.792</td>
<td>.752</td>
<td>.697</td>
<td>.548</td>
</tr>
<tr>
<td>Feldt-Gilmer</td>
<td>.800</td>
<td>.798</td>
<td>.786</td>
<td>.752</td>
<td>.696</td>
<td>.547</td>
</tr>
<tr>
<td>Lambda2</td>
<td>.800</td>
<td>.798</td>
<td>.782</td>
<td>.753</td>
<td>.702</td>
<td>.586</td>
</tr>
<tr>
<td>Maximized lambda4</td>
<td>.800</td>
<td>.798</td>
<td>.786</td>
<td>.781</td>
<td>.760</td>
<td>.703</td>
</tr>
<tr>
<td>Standardized alpha</td>
<td>.800</td>
<td>.799</td>
<td>.799</td>
<td>.753</td>
<td>.698</td>
<td>.549</td>
</tr>
<tr>
<td>Components 1 &amp; 2 vs. Components 3 &amp; 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.800</td>
<td>.798</td>
<td>.763</td>
<td>.694</td>
<td>.570</td>
<td>.234</td>
</tr>
<tr>
<td>Angoff-Feldt</td>
<td>.800</td>
<td>.799</td>
<td>.806</td>
<td>.695</td>
<td>.570</td>
<td>.235</td>
</tr>
<tr>
<td>Lambda2</td>
<td>.800</td>
<td>.798</td>
<td>.763</td>
<td>.694</td>
<td>.570</td>
<td>.234</td>
</tr>
<tr>
<td>Stratified alpha</td>
<td>.800</td>
<td>.798</td>
<td>.786</td>
<td>.781</td>
<td>.760</td>
<td>.703</td>
</tr>
<tr>
<td>Standardized alpha</td>
<td>.800</td>
<td>.799</td>
<td>.799</td>
<td>.694</td>
<td>.570</td>
<td>.235</td>
</tr>
<tr>
<td>Maximal reliability</td>
<td>.800</td>
<td>.801</td>
<td>.801</td>
<td>.784</td>
<td>.764</td>
<td>.708</td>
</tr>
<tr>
<td>Components 1 &amp; 2 vs. Component 3 vs. Component 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.750</td>
<td>.748</td>
<td>.673</td>
<td>.683</td>
<td>.605</td>
<td>.396</td>
</tr>
<tr>
<td>Raju</td>
<td>.800</td>
<td>.798</td>
<td>.718</td>
<td>.694</td>
<td>.615</td>
<td>.402</td>
</tr>
<tr>
<td>Feldt</td>
<td>.800</td>
<td>.808</td>
<td>.820</td>
<td>.738</td>
<td>.655</td>
<td>.430</td>
</tr>
<tr>
<td>Kristof</td>
<td>.800</td>
<td>.798</td>
<td>.786</td>
<td>.703</td>
<td>.608</td>
<td>.422</td>
</tr>
<tr>
<td>Lambda2</td>
<td>.760</td>
<td>.758</td>
<td>.690</td>
<td>.688</td>
<td>.606</td>
<td>.410</td>
</tr>
<tr>
<td>Standardized alpha</td>
<td>.787</td>
<td>.797</td>
<td>.797</td>
<td>.742</td>
<td>.677</td>
<td>.508</td>
</tr>
<tr>
<td>Maximal reliability</td>
<td>.800</td>
<td>.804</td>
<td>.804</td>
<td>.749</td>
<td>.688</td>
<td>.549</td>
</tr>
</tbody>
</table>

Note. To assess the degree of heterogeneity in the two-factor data, the observed covariances of the simulated data were factor analyzed using confirmatory factor analysis. The correlations among the factors were .80, .40, and .20, respectively, in slightly, moderately, and strongly heterogeneous data. Tau eq. = tau equivalent; congén. = congeneric equivalent.
work. However, the strongly heterogeneous condition was useful for determining whether alternatives to coefficient alpha could improve the estimate of reliability even in this extreme condition.

**Analysis of two halves: Two components each half.**

In the heterogeneous split-half simulations coefficient alpha strongly underestimated the true reliability (see Table 1). Coefficient alpha was 89% of true reliability when the components were slightly heterogeneous, 75% of the true reliability when the components were moderately heterogeneous, and only 33% of true reliability when the components were strongly heterogeneous. However, in this situation stratified alpha was exactly equal to the true reliability. Maximal reliability very slightly overestimated the true validity.

**Analysis of three parts: Part 1, Components 1—2; Part 2, Component 3; Part 4, Component 4.** As can be seen from Table 1, when the components were parallel in three-part simulations with an unequal number of components in the parts, only the Raju, Feldt, and Kristof coefficients and maximal reliability were equal to the true reliability. This is because the parts were only congeneric equivalent even though the components were parallel. When the components were tau equivalent, only Raju and Kristof coefficients were equal to the true reliability because the parts were only congeneric. The Feldt coefficient did not equal the true reliability because it assumes that the components are parallel. When the components were congeneric equivalent, only Kristof's coefficient was exactly equal to the true reliability. All coefficients underestimated true reliability when the components were generated by two factors. Coefficient alpha was about 87% of the true reliability when the components were slightly heterogeneous, and 80% and 56% of true reliability when the components were moderately and strongly heterogeneous, respectively. Maximal reliability was the best estimator for these data, but in the strongly heterogeneous condition maximal reliability was still only 78% of true reliability.

**Summary of simulations of four-component data.** In these simulations coefficient alpha was fairly robust in ungrouped single-component data. However, even in this condition coefficient alpha was only 92% of the true reliability when the data were moderately heterogeneous. Coefficient alpha computed on two-part splits when the components in each part loaded on separate factors tended to strongly underestimate the true reliability in the heterogeneous conditions, but stratified alpha was equal to the true reliability and maximal reliability overestimated the true reliability only very slightly in all such conditions. Maximized lambda4 was equal to true reliability in all simulations. Because with four components there are only three possible split halves, maximized lambda4 should routinely be computed when studying the reliability of the average of four components.

**Results for Eight-Component Data**

**Analysis of eight components ungrouped.** Results for the eight-component simulations are presented in Table 2. On the whole the results mirrored the four-component simulations, that is, all coefficients were equal to the true reliability when the components were parallel, all except standardized alpha were equal to the true reliability when the components were tau equivalent, and only the Feldt-Gilmer coefficient and maximized lambda4 equaled the true reliability when the components were congeneric equivalent. However, in these simulations coefficient alpha underestimated the true reliability only to a moderate degree. Coefficient alpha was 98% of the true reliability in slightly heterogeneous condition and 96% and 90% of the true reliability in the moderate and strongly heterogeneous conditions, respectively. Lambda2 was a slightly better estimator than coefficient alpha, but the differences were not large.

**Analysis of two halves: Four components each half.** As in the four-component data, coefficient alpha did not perform nearly as well in these simulations. Coefficient alpha was 89% of true reliability in the slightly heterogeneous case and 75% and 33% of true reliability in the moderately and strongly heterogeneous conditions, respectively. However, stratified alpha was exactly equal to the true reliability, and maximal reliability overestimated the true reliability only very slightly in the heterogeneous conditions. The other estimators except for maximized lambda4 were very close to coefficient alpha and did not improve the estimate to any great extent.

**Analysis of three parts: Part 1, Components 1—3; Part 2, Components 4—6; Part 3, Components 7—8.** Results on three-part data are presented in Table 2. Raju, Feldt, and Kristof coefficients and stratified alpha and maximal reliability were equal to the true reliability when the components were parallel, and only Raju, Kristof, and stratified alpha coefficients were equal to the true reliability when the components were tau equivalent, rendering the three parts only congeneric equivalent. Coefficient alpha was 95%, 95%, and 77%, respectively, of the true reliability in the slightly, moderately, and strongly
Table 2

<table>
<thead>
<tr>
<th>Degree of heterogeneity</th>
<th>Parallel</th>
<th>Tau eq.</th>
<th>Congen.</th>
<th>Slight</th>
<th>Moderate</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>True reliability</td>
<td>.727</td>
<td>.725</td>
<td>.712</td>
<td>.704</td>
<td>.679</td>
<td>.613</td>
</tr>
<tr>
<td>Eight components ungrouped</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.727</td>
<td>.725</td>
<td>.709</td>
<td>.692</td>
<td>.654</td>
<td>.554</td>
</tr>
<tr>
<td>Feldt</td>
<td>.727</td>
<td>.725</td>
<td>.715</td>
<td>.693</td>
<td>.654</td>
<td>.555</td>
</tr>
<tr>
<td>Feldt–Gilmer</td>
<td>.727</td>
<td>.725</td>
<td>.712</td>
<td>.692</td>
<td>.654</td>
<td>.554</td>
</tr>
<tr>
<td>Lambda2</td>
<td>.727</td>
<td>.725</td>
<td>.711</td>
<td>.693</td>
<td>.657</td>
<td>.571</td>
</tr>
<tr>
<td>Maximized lambda4</td>
<td>.727</td>
<td>.725</td>
<td>.712</td>
<td>.704</td>
<td>.679</td>
<td>.613</td>
</tr>
<tr>
<td>Standardized alpha</td>
<td>.727</td>
<td>.727</td>
<td>.727</td>
<td>.694</td>
<td>.656</td>
<td>.556</td>
</tr>
<tr>
<td>Components 1–4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs. Components 5–8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.727</td>
<td>.725</td>
<td>.691</td>
<td>.625</td>
<td>.599</td>
<td>.204</td>
</tr>
<tr>
<td>Angoff–Feldt</td>
<td>.727</td>
<td>.726</td>
<td>.735</td>
<td>.626</td>
<td>.599</td>
<td>.445</td>
</tr>
<tr>
<td>Lambda2</td>
<td>.727</td>
<td>.725</td>
<td>.691</td>
<td>.625</td>
<td>.599</td>
<td>.204</td>
</tr>
<tr>
<td>Stratified alpha</td>
<td>.727</td>
<td>.725</td>
<td>.712</td>
<td>.704</td>
<td>.679</td>
<td>.613</td>
</tr>
<tr>
<td>Standardized alpha</td>
<td>.727</td>
<td>.726</td>
<td>.726</td>
<td>.626</td>
<td>.599</td>
<td>.204</td>
</tr>
<tr>
<td>Maximal reliability</td>
<td>.727</td>
<td>.728</td>
<td>.728</td>
<td>.706</td>
<td>.679</td>
<td>.613</td>
</tr>
<tr>
<td>Components 1–3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs. Components 4–6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vs. Components 7 &amp; 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.716</td>
<td>.714</td>
<td>.683</td>
<td>.667</td>
<td>.612</td>
<td>.449</td>
</tr>
<tr>
<td>Raju</td>
<td>.727</td>
<td>.725</td>
<td>.694</td>
<td>.678</td>
<td>.622</td>
<td>.477</td>
</tr>
<tr>
<td>Feldt</td>
<td>.727</td>
<td>.728</td>
<td>.726</td>
<td>.681</td>
<td>.625</td>
<td>.480</td>
</tr>
<tr>
<td>Kristof</td>
<td>.727</td>
<td>.725</td>
<td>.712</td>
<td>.680</td>
<td>.633</td>
<td>.597</td>
</tr>
<tr>
<td>Lambda2</td>
<td>.721</td>
<td>.719</td>
<td>.695</td>
<td>.672</td>
<td>.618</td>
<td>.484</td>
</tr>
<tr>
<td>Stratified alpha</td>
<td>.727</td>
<td>.725</td>
<td>.711</td>
<td>.699</td>
<td>.668</td>
<td>.578</td>
</tr>
<tr>
<td>Standardized alpha</td>
<td>.723</td>
<td>.726</td>
<td>.723</td>
<td>.679</td>
<td>.625</td>
<td>.487</td>
</tr>
<tr>
<td>Maximal reliability</td>
<td>.727</td>
<td>.728</td>
<td>.728</td>
<td>.702</td>
<td>.674</td>
<td>.612</td>
</tr>
</tbody>
</table>

Note. To assess the degree of heterogeneity in the two-factor data, the observed covariances of the simulated data were factor analyzed using confirmatory factor analysis. The correlations among the factors were .80, .40, and .20, respectively, in slightly, moderately, and strongly heterogeneous data. Tau eq. = tau equivalent; cong. = congeneric equivalent.

Summary of simulations for eight-component data.

Coefficient alpha performed reasonably well in the eight-part simulations, equaling 95% of true reliability in moderately heterogeneous data. When the components were grouped into two parts with separate factors in each part, coefficient alpha was only 75% of true reliability in moderately heterogeneous data. However, in this condition, stratified alpha was exactly equal to the true reliability and maximal reliability overestimated the true reliability only by a very small margin. It is clear that coefficient alpha computed on individual components was a better lower bound than that computed on components grouped into two halves that were not highly correlated. Maximized lambda4 was equal to the true reliability in all simulations.

Implications for Corrections for Attenuation

In four-component simulations with moderate heterogeneity, coefficient alpha ranged from 75% to 92% of true reliability. Under these conditions, using coefficient alpha in single-variable corrections for attenuation would result in overcorrections of 4% to 15%. For example, if true validity were actually .50, estimated true validity would range from .52 to .58. In the worst case scenario, when the components in each part loaded on separate factors and the data were strongly heterogeneous, coefficient alpha was only 33% of true reliability, and if used in the correction for attenuation would overestimate true reliability by 74%.
In eight-component simulations with moderate heterogeneity, coefficient alpha ranged from 75% to 96% of true reliability. Under these conditions, using coefficient alpha in corrections for attenuation in one variable could result in an overcorrection of 2% to 15%. For example, if the true validity was actually .50, estimated true validity would range from .51 to .58. In the worst-case scenario, in which the data were strongly heterogeneous and the components in each split loaded on separate factors, coefficient alpha was only 33% of true reliability. In this extreme case, if the true validity were actually .50, estimated true validity would be .87.

Finally, if coefficient alpha was used to correct both independent and dependent variables for unreliability and if both variables were moderately heterogeneous, in the worst-case scenario the double correction would overestimate the true validity by about 33%.

Discussion and Recommendations

One-Factor Models

When the components of a measure were parallel and each part contained an equal number of components, all estimators were equal to the true reliability. However, this condition is seldom realized in real data because it means that the components (items, subtests, etc.) have equal factor loadings and equal error variances. When the parts contained an unequal number of parallel components, the parts were only congeneric equivalent, and only Raju, Feldt, and Kristof coefficients and stratified alpha and maximal reliability estimates equaled the true reliability.

When the components were tau equivalent and there were an equal number of components in each part, all estimates were equal to the true reliability except Angoff–Feldt, standardized alpha, and maximal reliability, which overestimated true reliability by a very small amount.

When the components were congeneric equivalent, only Feldt–Gilmer, maximized lambda4, stratified alpha, and Kristof coefficients were equal to the true reliability. However, when the data were congeneric, almost all coefficients could be used as adequate estimates of the true reliability. The exception was coefficient alpha computed on three-part splits with an unequal number of components in the splits.

Two-Factor Models

When the data were generated by two underlying factors and the components in one subset measured one factor and the components in the other subset measured a different factor, coefficient alpha seriously underestimated the true reliability. Lambda2 tended to be slightly larger than coefficient alpha but did not improve the estimate of true reliability to any great extent. However, in these conditions stratified alpha and maximal reliability both provided adequate estimates of the true reliability.

Maximized lambda4 was clearly the most consistently accurate estimator of the true reliability in all the simulations that were studied. For measures with few components, if one suspects that data are heterogeneous, lambda4 should be computed for all possible splits and maximized lambda4 used to estimate the true reliability. For measures with 6 to 10 components, the situation is problematical. The approximation to maximized lambda4 presented in this article could be computed and used to estimate true reliability. However, maximized lambda4 may slightly overestimate the true reliability because of capitalizing on sampling errors.

Assessing Heterogeneity

Unfortunately, the characteristics of empirical data (tau equivalent, congeneric, or multifactorial) are seldom known precisely. However, confirmatory factor analysis (CFA) is useful for modeling and distinguishing between parallel, tau-equivalent, and congeneric data providing that the assumptions of this approach are reasonably approximated (see Turban, Sanders, Frances, & Osburn, 1989). Another good reason for using CFA is the possibility of correlated errors among the components. Komaroff (1997) has shown that correlated errors can inflate coefficient alpha and that such errors interact with lack of tau equivalence. Correlated errors can be detected with CFA, and if necessary alpha can be adjusted for this effect. Also, there has been a good deal of work on ways to determine whether data is uni- or multidimensional, and a large number of indices have been proposed (see Hattie, 1985). Factor analysis (either exploratory or confirmatory) of the components is recommended to determine the heterogeneity of the data. If the data are not continuous, it is also important to factor the polyphoric correlations among the components. If the data set is characterized by two or more moderately correlated underlying factors (in the range of .20 to .40 or so), an investigator should be concerned. The exception to this rule is the situation in which there are multiple factors but the first factor is dominant in the
sense of accounting for the bulk of the variance among the components (Cronbach, 1951).

It is important to note that the average intercomponent correlation is not a reliable indicator of heterogeneity. The average correlation among components may be low, but the data may still be tau equivalent, and therefore true reliability is accurately estimated by coefficient alpha (Green, Lissitz, & Mulaik, 1977). Variations in the correlations among components may also be a somewhat misleading indicator of heterogeneity. Even if there are variations in intercomponent correlations, the data may still be tau equivalent. However, intercomponent correlations that cluster into subgroups indicate heterogeneity (Cortina, 1993).

If the data are very heterogeneous, components can be grouped on the basis of similar content, and stratified alpha or maximal reliability along with coefficient alpha should be computed and the largest value used to estimate the true reliability. If there is no basis for grouping and there are a small number of components, lambda4 should be computed for all possible splits.

**Many Versus Few Components**

This study assessed coefficient alpha computed on only four and eight components. The reason for studying such a small number of components is that coefficient alpha is more often than not computed on eight or fewer components (Peterson, 1994) and because coefficient alpha approaches the true reliability as the number of positively correlated components increases. Coefficient alpha becomes less and less of an underestimate of the true reliability because the percent increase in coefficient alpha as the number of components increases is greater than the percent increase in true reliability. For example, in the present study, to limit overestimation of corrected validity to about 2%, coefficient alpha must be at least 95% of the true reliability. In the worst-case scenario (strong heterogeneity) for eight components ungrouped, coefficient alpha was 90% of true reliability, but in this case only 20 components would be needed for coefficient alpha to reach 95% of true reliability. In the worst-case scenario for four components ungrouped, coefficient alpha was 78% of true reliability, but in this case only 28 components would be needed for coefficient alpha to reach 95% of true reliability. It is concluded that when estimating the reliability of questionnaire or rating data, even if the data are very heterogeneous, only 20 to 30 components are needed to be reasonably assured of an adequate estimate of true reliability. However, in the case of grouped data increasing, the number of components in each group does not necessarily make coefficient alpha a better lower bound for alternative estimates of the true reliability to be used. For example, in the worst-case scenario for two groups of four components in each group, coefficient alpha was only 33% of true reliability. However, in this case one can compute stratified alpha, which is 100% of true reliability.

**References**


---

**Appendix**

**Simulation Method**

1. Tau-equivalent data were simulated as follows:
   a. The correlations among the component true scores were set to 1.0.
   b. The variances of the component true scores were set to 1.0.

   These two operations generate a true-score covariance matrix with every element unity. The matrix is 4 x 4 for four-component data and 8 x 8 for eight-component data.

c. To generate the corresponding tau-equivalent observation matrix, the reliabilities of the components were systematically varied. For example, in the four-component case the reliabilities were set to .45, .46, .52, and .55. The observed-score correlation between the ith and jth component observed scores was computed by multiplying the true-score correlation between the ith and jth components by the square root of the product of their respective reliabilities. For example, in the four-component case the observed-score correlation between Components 1 and 2 (.465) was computed by multiplying 1.0 by the square root of .45 x .48.

d. Because reliability is true-score variance divided by observed-score variance, the observed-score variance of the ith component was computed by dividing the true variance of the component by its respective reliability. For example, the observed variance of Component 1 in the four-component case (2.15) was computed by dividing 1.0 by .465.

2. Parallel data were simulated the same way as for tau-equivalent data except that the reliabilities of the components were kept constant.

3. Congenetic-equivalent data were simulated the same way as for tau-equivalent data except that the reliabilities of the components were systematically varied.

4. True reliability was computed by dividing the variance of the summed true score on the components by the variance of the summed observed score.
5. Supervisor-rating data with four components were simulated as follows:

a. To simulate parallel data, the reliability of each component was set at .50. This resulted in parallel data with correlations of .50 among the observed scores.

b. To simulate tau-equivalent data, the reliabilities of the components were set to .45, .48, .52, and .55. This resulted in tau-equivalent data with correlations that were variable (.465 to .535) among the components but averaging approximately .50.

c. To simulate congeneric-equivalent data, the true variance of Components 1 and 2 was set at 2.0 and the true variance of Components 3 and 4 was set at 1.0.

d. To generate slightly heterogeneous data, true-score correlations between Components 1 and 2 with Components 3 and 4 were set to .8; for moderately heterogeneous data 0.80s were replaced by 0.60s. For strongly heterogeneous data 0.60s were replaced by 0.20s. Confirmatory factor analysis (CFA) applied to all three observed covariance matrices resulted in a two-factor solution. Components 1 and 2 loaded 1.00 on Factor 1 and 0 on Factor 2. Components 3 and 4 loaded 0 on Factor 1 and 1.00 on Factor 2. For slightly heterogeneous data, the correlation between Factors 1 and 2 was .80. For moderately heterogeneous data the correlation was .60, and for strongly heterogeneous data the correlation was .20.

e. Four-part data were simulated by $X = X_1 + X_2 + X_3 + X_4$; two-part data were simulated by $X_a = X_1 + X_2$, $X_b = X_3 + X_4$; three-part data were simulated by $X_a = X_1 + X_2$, $X_b = X_3$, $X_c = X_4$.

6. Questionnaire data with eight components were simulated as follows:

a. To simulate parallel data, the reliability of each component was set at .25. This resulted in correlations of .25 among the observed scores.

b. To simulate tau-equivalent data, the reliabilities of the components were set to .22, .24, .24, .26, .26, .28, and .28. This produced tau-equivalent data with observed score correlations ranging from .22 to .28 but averaging approximately .25.

c. To simulate congeneric-equivalent data, the true variances were set to 2.0 for Components 1, 2, 3, and 4, and to 1.0 for Components 5, 6, 7 and 8.

d. To generate slightly heterogeneous data, true-score correlations between Components 1, 2, 3, and 4 with Components 5, 6, 7, and 8 were set to .8, for moderately heterogeneous data 0.80s were replaced by 0.60s. For strongly heterogeneous data 0.60s were replaced by 0.20s. These data were factored using CFA. The factor loadings were 1.0 and 0.0, and correlations among the two factors were the same as for four-component data, that is, .8, .6, and .2.

e. Eight-part data were simulated by $X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8$; two-part data were simulated by $X_a = X_1 + X_2 + X_3 + X_4$, $X_b = X_5 + X_6 + X_7 + X_8$; three-part data were simulated by $X_a = X_1 + X_2 + X_3$, $X_b = X_4 + X_5 + X_6$, $X_c = X_7 + X_8$.

Received October 11, 1996
Revision received July 20, 1999
Accepted March 22, 2000