STRUCTURAL EQUATION MODELING, 2(3), 255–273

TEACHER'S CORNER

Coefficient Alpha: A Basic Introduction From the Perspectives of Classical Test Theory and Structural Equation Modeling

Michael B. Miller
Washington University School of Medicine

This article is a pedagogical piece on coefficient alpha (α) and its uses. The classical approach to test reliability is explained. Test-retest, alternative-forms, and internal-consistency methods of approximating test reliability are described, equations are derived for each method, and α is shown to be a lower-bound internal-consistency approximation to test reliability. Emphasis is placed on the effects of violations of model assumptions on reliability estimation. The classical models are conceptualized as structural equation models and are displayed in path diagrams. Special emphasis is placed on the failure of α to meet certain basic criteria as an index of test homogeneity.

Coefficient alpha (α) is a commonly used index of test reliability. It can be used for any test on which scores are produced by summing the scores of two or more test items. It was developed by several researchers during the first half of this century and was named by Cronbach (1951), the first writer to recognize its general usefulness. It is often called *Cronbach's alpha* in honor of his contribution. Unfortunately, practitioners have not always understood α and some have misinterpreted it as a measure of test homogeneity or unidimensionality. This article uses concepts from classical test theory and

Requests for reprints should be sent to Michael B. Miller, Department of Psychiatry, Medical School Box 8134, Washington University School of Medicine, 4940 Children's Place, St. Louis, MO 63110.

modern structural equation modeling (SEM) to explain the meaning of α and its proper and improper uses. The discussion is meant to be accessible to students with little knowledge of classical test theory and SEM, but references are given for further reading at a much more advanced level. Throughout the article I will refer to tests measuring "traits of examinees." I do this for the sake of clarity and consistency, and I hope it doesn't obscure the generality of the methods and concepts. Readers should feel free to translate the trait of an examinee into the languages of their own fields, for example, "social skill of a child" or "productivity of a farm," or "severity of illness of a cancer patient."

CLASSICAL CONCEPT OF TEST RELIABILITY

Imagine that a test is administered to every examinee in some defined population of examinees and scores are obtained. Part of the variation in scores among examinees will be due to genuine or "true" underlying differences among the examinees in the trait being measured. Of course, no test is perfect and we expect that an additional part of the variation in scores will be due to random errors of measurement. In classical test theory the measurement error is defined to be uncorrelated with the genuine or true contribution to the score. This implies that we can write the variance of the test scores as the sum of the true-score variance plus the error variance. Test score reliability is a variance ratio equal to the true-score variance of the test scores divided by the total variance of the test scores. Keep in mind that the reliability of a test is not a property of a test per se; rather, the reliability is a property of a test administered to a particular population of examinees under certain conditions. Test scores will have a higher reliability when there is a larger genuine variation in the population of examinees. For example, a math test constructed for third graders will show a higher reliability when administered to all of the students in Grades 1 through 5 than when administered to Grade 3 alone.

Definition of Test Reliability

In classical test theory (see Lord & Novick, 1968, for a formal mathematical account or Nunnally, 1978, for something less mathematical) an observed test score (X) is composed of two latent independent components, a true score (T) and an error score (E) such that

$$X = T + E \tag{1}$$

Figure 1a shows a path diagram representation of Equation 1. In this style of path diagram a curved arrow from an exogenous (upstream) variable to

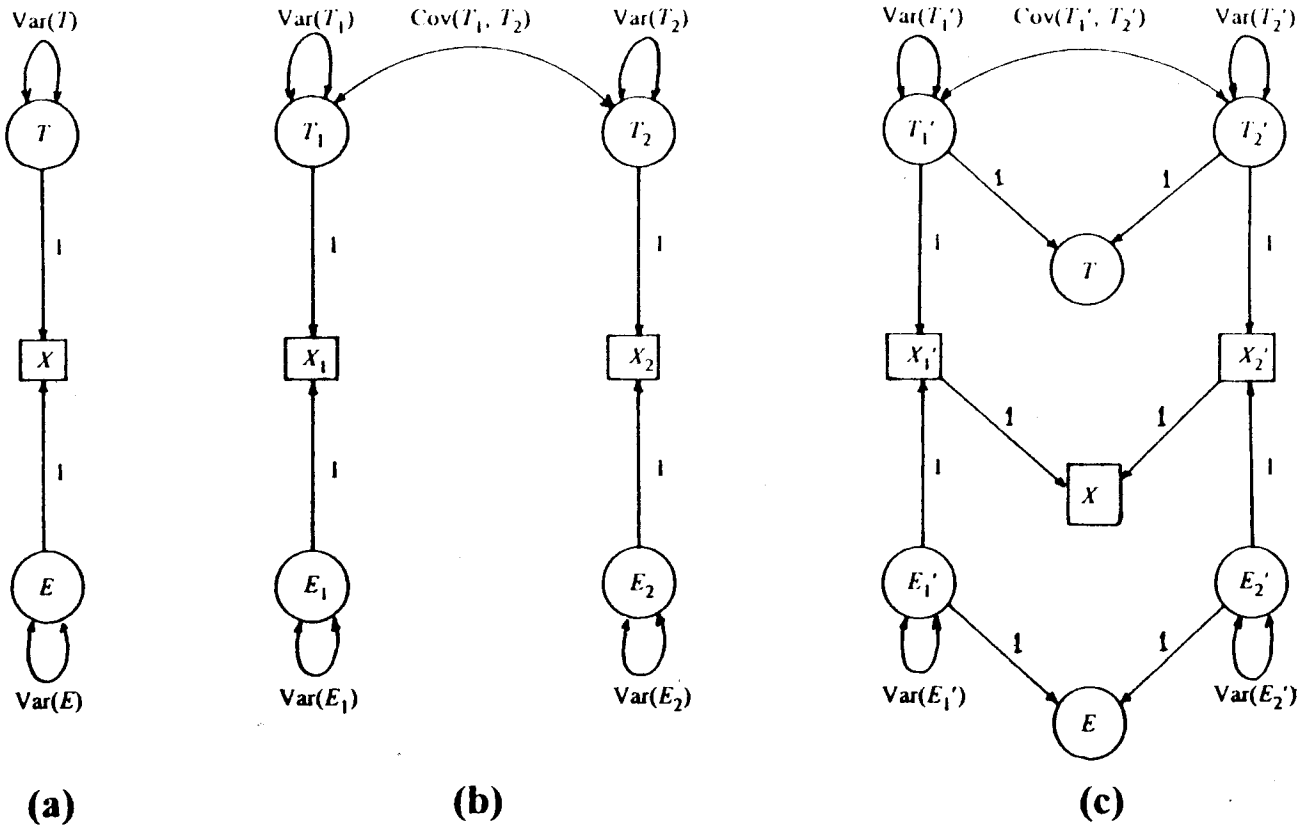

FIGURE 1 (a) Path diagram representing the classical model expressed in Equations 1 and 2. (b) Path diagram for test-retest or alternative-forms data as represented by Equations 4a and 4b. (c) Path diagram for split-halves of a single test.

itself represents the variance of that variable (McArdle & Boker, 1990; Neale & Cardon, 1992). We can write the variance of X in terms of the variances of T and E.

$$\text{Var}(X) = \text{Var}(T) + \text{Var}(E) \tag{2}$$

Equation 2 shows the covariance structure for the model shown in Figure 1a. The model is underidentified—there are two model parameters, the variances of T and E, but only one statistic, the variance of X. The reliability of X, represented by ρ_{xx}, can be expressed as

$$\rho_{xx} = \text{Var}(T)/\text{Var}(X) \tag{3}$$

With only one score per examinee, the model is underidentified and it is impossible to estimate reliability.

Notation. The classical test theory concepts just presented can be used in the measurement model part of a structural equations model. In LISREL notation form, ξ can be used instead of T and δ instead of E. Not everyone agrees that classical and SEM variables should be equated in this way (see Bollen, 1989, p. 219) because δ can include a specific, nonerror component. For consistency, I will continue to use the T and E notation.

APPROXIMATION OF TEST RELIABILITY: PERSPECTIVES FROM CLASSICAL TEST THEORY

Two limitations make it impossible to know the exact value of a test's reliability: (a) true scores and error scores can't be observed directly, so the appropriate model for their relations is always unknown, and (b) population parameters can only be estimated from sample data. This section contends primarily with the first of these problems. The classical models make explicit assumptions about the relations of latent true and error scores so that equations can be produced in which test reliability is written in terms of variances and covariances of observed scores. I will refer to these classical reliability coefficients as *approximations* of true test reliability.

Many methods are available for approximation of test reliability, but they can be grouped into three general categories: test-retest methods, alternative-forms methods, and internal-consistency methods. These methods approach approximation in different ways and the reliability coefficients they yield should be interpreted in different ways. Cronbach (1947) wrote a classic article on the subject, and Bollen's (1989) discussion of this and related matters will be of interest to most readers.

Suppose that two tests (Test 1 and Test 2) are available for the measurement of some trait. We will call the scores from these tests (for an examinee selected at random from a given population) X_1 and X_2, and we'll assume that these test scores are composed of latent true-score and error-score components,

$$X_1 = T_1 + E_1, \text{ and} \tag{4a}$$

$$X_2 = T_2 + E_2 \tag{4b}$$

Still following the classical theory, we will assume that the error scores are uncorrelated both with other error scores and with the true scores. Thus,

$$\text{Cov}(E_1, E_2) = \text{Cov}(E_1, T_1) = \text{Cov}(E_2, T_2) = \text{Cov}(E_1, T_2) = \text{Cov}(E_2, T_1) = 0$$

This model is the basic foundation from which all three types of reliability approximations are derived.

Test-Retest and Alternative-Forms Methods

The path diagram shown in Figure 1b represents Equations 4a and 4b, which apply both to test-retest data and to alternative-forms data. The diagram also depicts the classical assumption that error scores are uncorrelated both with true scores and with other error scores. In the test-retest method, we define scores X_1 and X_2 to be scores from the same test administered to the same person at Time 1 and Time 2. The alternative-forms method is identical to the test-retest method except that an alternative form of the test is administered at Time In both methods, the correlation of X_1 with X_2, $\text{Corr}(X_1, X_2)$, can be used to approximate the reliability (ρ_{xx}) of the test. If we assume that the variance of the test is the same at the two points in time, $\text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(X)$, we see that $\text{Var}(X) = \sqrt{\text{Var}(X_1) \cdot \text{Var}(X_2)}$. Using the usual formula for a covariance of sums and the assumptions of the classical model is,

$$\text{Cov}(X_1, X_2) = \text{Cov}(T_1, T_2) \tag{5}$$

The same result can be derived from the path diagram simply by tracing the path from X_1 to X_2. This model is underidentified. There are five parameters (four variances and a covariance), but only three statistics (two variances and a covariance). We can't estimate any variance parameters unless we make a restrictive assumption.

Essential tau-equivalence. We can produce an approximation of true-score variance using the restriction that $T_1 = T_2 + c = T$, where c is an additive

constant that shifts the true score up or down by the same amount for every examinee. This assumption, that true scores from two tests differ only by the same additive constant for every examinee, is called *essential tau-equivalence*. Essential tau-equivalence implies that $\mathrm{Var}(T_1) = \mathrm{Var}(T_2) = \mathrm{Var}(T)$, and that the correlation of true scores from the two testings is perfect, that is, $\mathrm{Corr}(T_1, T_2) = 1.0$. We can now rewrite the covariance of true scores under this restrictive assumption.

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(T_1, T_2) = \mathrm{Corr}(T_1, T_2)\sqrt{\mathrm{Var}(T_1) \cdot \mathrm{Var}(T_2)} = \mathrm{Var}(T) \qquad (6)$$

Thus, combining information from Equation 6 with the definition of a correlation coefficient, we see that under the stated assumptions,

$$\mathrm{Corr}(X_1, X_2) = \mathrm{Cov}(X_1, X_2)/\sqrt{\mathrm{Var}(X_1) \cdot \mathrm{Var}(X_2)} = \mathrm{Var}(T)/\mathrm{Var}(X) = \rho_{xx}$$

The assumption of essential tau-equivalence constrains the true-score variances to equal the covariance of true scores, $\mathrm{Var}(T_1) = \mathrm{Var}(T_2) = \mathrm{Cov}(T_1, T_2)$, and makes the model identified. Note that we have obtained a single reliability estimate for both X_1 and X_2 by assuming that they have equal variance and equal true-score variance. In a SEM framework, variance parameters could be estimated under the same restrictions and the reliability could then be estimated as $\rho_{xx} = \mathrm{Var}(T_1)/[\mathrm{Var}(T_1) + \mathrm{Var}(E_1)]$. The advantage of using the sample correlation coefficient in practice is that it is much easier to obtain and it will always yield a very similar result, in fact, the result will be identical when the sample variances for X_1 and X_2 are equal.

What if these assumptions are violated? If the assumption of essential tau-equivalence alone is violated, $\mathrm{Cov}(T_1, T_2)$ will be less than $\mathrm{Var}(T)$ and $\mathrm{Corr}(X_1, X_2)$ will be smaller than ρ_{xx}. But, violations of other assumptions may cause the test-retest approximation to be greater than test reliability. Specifically, the covariance of errors from Test 1 and Test 2 is likely to be positive because examinees may tend to repeat at Time 2 the answers they remember giving at Time 1. Because of this, the test-retest approximation is usually considered to be larger than test reliability. The assumption that $\mathrm{Cov}(E_1, E_2) = 0$ is much more reasonable with alternative forms—different items are administered at the two testings, so it is not possible for an examinee to remember an item from an earlier session. However, the assumption of essential tau-equivalence appears more likely to be violated when alternative forms are administered. Thus, the alternative-forms method tends to give a lower approximation of reliability than that obtained by the test-retest method. Of course, in choosing a method of reliability estimation one must take into consideration the considerable time and expense involved in the development of an alternative form.

Internal-Consistency Methods

The most popular internal-consistency methods of reliability estimation are the split-half method and coefficient alpha. These methods can be used only when a test score is a sum of component scores. Component scores could be individual item scores or weighted item scores or subtest scores. The internal-consistency methods of reliability estimation to be considered here assume that the test is completely homogeneous, meaning that all components load on a single common true-score factor and all unique variance is measurement error. The assumption of a single common factor is reasonable when items have been selected carefully to measure a single trait dimension. Internal-consistency approaches to reliability are sensitive to test homogeneity so that, other things being equal (and they never are), tests that are less homogeneous will yield lower internal-consistency reliability approximations.

Effect of test length on test reliability. A longer test is generally more reliable than a shorter one. Intuitively, we know that a longer test yields more information and should therefore be more reliable. Whether this relation holds in reality depends on whether the additional items are correlated positively with the items already present. If the intercorrelations of the new items are like those of the items already present, the reliability of the longer test can be predicted from the reliability of the shorter test using the Spearman–Brown "prophecy formula" (see Nunnally, 1978). The issue of test length is presented at this point in the article because it is relevant only to tests that are composed of components or items. However, test length affects not only internal-consistency estimates of reliability, but all other types of estimates as well.

Split-half reliability estimates. In the split-half method of reliability estimation, a test is split into two halves of equal length. That is, half the test items are assigned to Subtest 1 and half to Subtest 2, either at random, or by putting odd-numbered items in one subtest and even-numbered items in the other. The correlation or covariance of these halves is computed and used in some equation (there are several from which to choose) to estimate the reliability of the whole test. For example, the correlation of two half-tests produces a direct estimate of the alternative-forms approximation of reliability for a test of half-length, which must be "stepped up," usually with the Spearman–Brown prophecy formula, to estimate the reliability for a test of full length. The split-half method has always been criticized (since at least the 1930s) because of the random or arbitrary nature of the split—different splits will produce different reliability estimates in practice. Which split should one use?

Derivation of Rulon's split-half method: Alpha for two items. We will now develop the mathematical rationale for one of the internal-consistency methods that can be used with split halves, the method of Rulon (1939). This method is important because it turns out to be equal to α for a 2-component (e.g., 2-item) test. Suppose that the observed test-score variable X is a composite of two observed component scores, X'_1 and X'_2, so that $X = X'_1 + X'_2$. These components can be partitioned into latent true-score and error-score components so that $X'_1 = T'_1 + E'_1$ and $X'_2 = T'_2 + E'_2$, with $T = T'_1 + T'_2$ and $E = E'_1 + E'_2$. (The "primes" [′] are used to remind us that these components are half as long as the whole test of interest, in contrast to the unprimed variables, X_1 and X_2, of the previous sections, which were equal in length to the whole test.) We assume that the classical test theory applies to these components so that errors are uncorrelated with each other and with the true scores. The path diagram in Figure 1c depicts all the equations showing the additive relations among the true, error, and observed scores of tests of both half- and full-length. Like all the models of the previous sections, this model is not identified. How can we use the component data to approximate the true-score variance of the full test? If we assume essential tau-equivalence, the model is identified, and it is not necessary to assume equal error variances for the two half tests before a reliability estimate for X can be computed. By tracing paths, or by the rule for variance of a sum, we know that $\mathrm{Var}(T) = \mathrm{Var}(T'_1) + \mathrm{Var}(T'_2) + 2\mathrm{Cov}(T'_1, T'_2)$, which reduces under essential tau-equivalence to $\mathrm{Var}(T) = 4\mathrm{Var}(T'_i) = 4\mathrm{Cov}(X'_1, X'_2)$, where subscript i represents either 1 or 2. Thus,

$$\rho_{xx} = 4\mathrm{Cov}(X'_1, X'_2)/\mathrm{Var}(X) \tag{7}$$

Cronbach (1951) demonstrated the equivalence of Rulon's method to those of several other writers and argued for its superiority over the aforementioned method of Spearman and Brown. As I mentioned earlier, application of Rulon's method to a 2-item test is identical to computing α for the same test. Instead of interpreting and critiquing Rulon's method at this point, I will derive α, the more general formula, and critique that.

Coefficient alpha (α). The derivation of α as an index of test reliability is a generalization of the derivation just shown of Rulon's coefficient for a two-component test. The difference is that α can be applied to tests with any number of items or components. Suppose our test score variable X is a sum of k item scores, $x_1, \ldots, x_k$. (See Figure 2a for a path diagram of a 4-item test.) The item scores can be expressed as sums of latent true and error components, $x_1 = t_1 + e_1, \ldots, x_k = t_k + e_k$, and the true and error components each

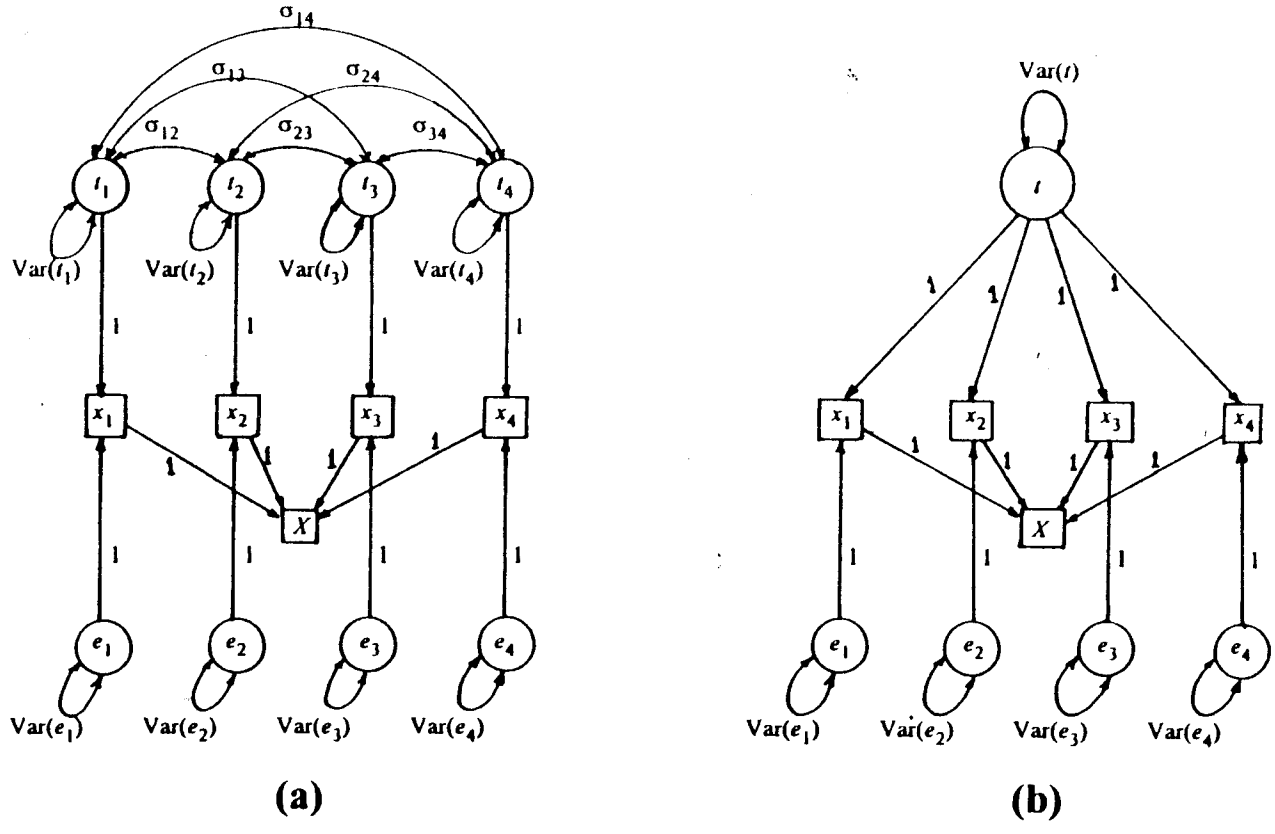

FIGURE 2 (a) General model for a test of four items. $Cov(t_i, t_j)$ is represented by σ_{ij}. (b) The same model under the constraint of essential tau-equivalence (i.e., $\sigma_{ij} = Var(t)$ for all i, j)

sum to the respective components of the test score, that is, $T = t_1 + t_2 + \ldots + t_k$, and $E = e_1 + e_2 + \ldots + e_k$. The model is obviously underidentified. Once again we will assume essential tau-equivalence of test items in order to produce an approximation of the test's true-score variance. (See Figure 2b for a path diagram of this model for four items.) Note that every interitem covariance and every item true-score variance is equal to $\mathrm{Var}(t)$ in this model. The error variances are not constrained. This model is overidentified so parameters can be estimated and the model can be tested, but one should bear in mind that terms designated as error scores will actually absorb the sum of specific item variance plus item error variance. This means that α may dramatically underestimate reliability even when the model fits the data very well.

We will now derive α from the path diagrams. First, compute the true-score variance of X from the path diagram in Figure 2b by counting all possible paths from X to t and back. Do the same thing with Figure 2a, always tracing back from X to either a variance or a covariance of a t_i variable. You should find a total of k^2 paths within each diagram such that the true-score variance of X can be written as $\mathrm{Var}(T) = k^2 \cdot \mathrm{Var}(t)$ for Figure 2b or $\mathrm{Var}(T) = \Sigma_{i=1}^{k}\mathrm{Var}(t_i) + 2\Sigma_{i>j}\sigma_{ij}$ for Figure 2a. The sum of covariances is doubled because each covariance is traced twice, once from t_i to t_j, and once from t_j to t_i. Under essential tau-equivalence, any interitem covariance equals $\mathrm{Var}(t)$, but for practical purposes it is much more sensible to approximate $\mathrm{Var}(t)$ using an average of all $k \cdot (k - 1) / 2$ distinct interitem covariances. Thus, $\mathrm{Var}(T) = k^2 \cdot \mathrm{Var}(t) = k^2 \cdot \mathrm{Ave}[\mathrm{Cov}(x_i, x_j)] = k^2 \cdot \Sigma_{i>j}\mathrm{Cov}(x_i, x_j) / [k \cdot (k - 1) / 2] = 2 \cdot [k/(k - 1)] \cdot \Sigma_{i>j}\mathrm{Cov}(x_i, x_j)$. The test variance is $\mathrm{Var}(X) = \Sigma_{i=1}^{k}\mathrm{Var}(x_i) + 2\Sigma_{i>j}\mathrm{Cov}(x_i, x_j)$, so that $2 \cdot \Sigma_{i>j}\mathrm{Cov}(x_i, x_j) = \mathrm{Var}(X) - \Sigma_{i=1}^{k}\mathrm{Var}(x_i)$, and $\mathrm{Var}(T) = [k/(k - 1)] \cdot [\mathrm{Var}(X) - \Sigma_{i=1}^{k}\mathrm{Var}(x_i)]$. We can now write the reliability as

$$\alpha = \rho_{xx} = \mathrm{Var}(T)/\mathrm{Var}(X) = [k/(k - 1)] \cdot [1 - \Sigma_{i=1}^{k}\mathrm{Var}(x_i)/\mathrm{Var}(X)] \qquad (8)$$

in the classical model under the highly restrictive assumption of essential tau-equivalence of every pair of items.

As a concrete example, suppose our four-item test has the following variance-covariance matrix of the four items.

	#1	#2	#3	#4
Item #1	16	5	8	2
Item #2	5	36	17	9
Item #3	8	17	49	7
Item #4	2	9	7	25

The sum of item variances is the trace of the matrix, that is, the sum of diagonal elements from upper left to lower right, and it equals 126. The

overall test variance is the sum of all elements of the matrix, and it equals 222. α can then be computed by Equation 8 as $(4 / 3) \times (1 - 126 / 222)$ or .58. An alternative approach is to estimate the item true-score variance, Var(t), as the average of the off-diagonal elements of the matrix. Then α can be computed as $k^2 \cdot \text{Var}(t)/\text{Var}(X) = 16 \times 8/222$ or .58.

The Appendix contains a LISREL (Jöreskog & Sörbom, 1994) script for the model of Figure 2b (leaving X out) that uses unweighted least squares to estimate Var(t) as the mean interitem covariance so that α can be computed as $k^2 \cdot \text{Var}(t)/\text{Var}(X)$. An Mx (Neale, 1994) script, also in the Appendix, computes α in a more direct fashion from the same covariance matrix.

INTERPRETATION AND CRITIQUE OF α

We are now in a position to discuss α, its meaning, and proper uses. α's lack of merit as an index of homogeneity or unidimensionality is discussed in detail in a later section.

Violations of Assumptions

Essential tau-equivalence and test homogeneity. We derived α as a reliability coefficient but noted that it would only equal the test reliability under the assumption that all pairs of items are essentially tau-equivalent. This assumption is mathematically identical to the assumption that all items have equal loadings on a single common factor with their unique variances composed entirely of error. A common factor model seems reasonable for many tests, especially when they were designed to be relatively homogeneous. The assumption of equal loadings seems less reasonable, especially when components have very different variances. What happens if these assumptions are violated? Novick and Lewis (1967), in an elegant analysis of the problem, proved that $\alpha \leq \rho_{xx'}$, with equality holding only under essential tau-equivalence. That is, α tends to underestimate test reliability. For an extreme example, imagine a very strange test in which every item has no error variance and true scores of all pairs of items are uncorrelated (instead of being correlated 1.0 as they would be under essential tau-equivalence). For such a test $\rho_{xx'} = 1.0$ but $\alpha = 0.0$. Keep this possibility in mind, but don't let it bother you too much! As Cronbach (1951) pointed out in response to Wherry and Gaylord's (1943) critique of internal-consistency estimates of reliability, tests with several uncorrelated factors (in the example just given there were k uncorrelated factors—one per item) do not yield interpretable scores. Thus, one might argue that for many tests, α is of greater interest than $\rho_{xx'}$ because low values of α warn us that either our test scores are not homogeneous, or our test scores are not reliable. On the other hand, α will still be less than 1.0 for a perfectly reliable and homogeneous test if factor

loadings are unequal. We should conclude then that α is a lower-bound approximation to test reliability even for a perfectly homogeneous test.

Uncorrelated errors. The classical model used to derive alpha assumed that error scores of all pairs of items are uncorrelated. This assumption was important because it allowed us to assume that the covariances of item scores were equal to the corresponding covariances of item true scores. If this classical assumption were violated, some interitem covariances would be increased so that they would be greater than the corresponding true-score covariances, and the approximation to test true-score variance would be too large. Thus, α would be inflated somewhat, but perhaps not beyond the true value of ρ_{xx}.

Sampling Variation of α

Until this point α and ρ_{xx} have been treated as population parameters. In reality, of course, these parameters are estimated from randomly sampled data and the estimates vary, sometimes markedly, from sample to sample. As always, this variation is decreased when samples are large. Samples are usually large in applications of SEM, so sampling variation will not often be a problem. It should be kept in mind, however, that variation due to sampling might easily cause estimated α (α_{est}) to exceed ρ_{xx}. Thus, our confidence in α_{est} as a lower bound to ρ_{xx} should be directly related to the size of the sample from which α_{est} is computed. Cortina (1993) correctly pointed out that the standard error of α will decrease with increasing number of test items. However, Cortina's standard error formula (his Formula 4, for which he cites no source) is clearly wrong because it fails to take sample size (number of examinees) into account. Reasonable formulas do exist for the computation of standard errors and for testing the hypothesis that α equals some particular value, but the validity of these formulas is dependent on distributional assumptions that will not often be met. (See Feldt, 1980; Feldt, Woodruff, & Salih, 1987; Pandey & Hubert, 1975).

α as the Average of All Split-Half Reliabilities

A problem for the split-half method of reliability estimation is that different splits give different results. The prospect of taking many splits and averaging the resulting reliability estimates is unappealing because of the cost in time and energy. Amazingly, Cronbach (1951) was able to prove that α is exactly equal to the average of all possible split-half reliabilities when the Rulon method is used. This holds true both for population α and for sample α_{est}. Therefore, there is little reason to use split-half estimates of reliability when the more robust α is so readily available.

α IS NOT AN INDEX OF TEST HOMOGENEITY

Researchers generally want the items of their multiple-item tests to "measure the same thing" or to "load on a single common factor." These concerns are stated imprecisely, but it is easy to see that the issue is important. We want to produce tests to measure particular traits and we want every item focused on measuring that same thing. If it should turn out that some items measure different things and show no relation to the other items of the test, we would want to eliminate them. If we had clusters of intercorrelated items that were not positively correlated with items in other clusters, we might choose to develop several tests to measure various aspects of the complex trait we'd started to measure with a single test. Some researchers have made the mistake of using α as an index of the homogeneity or unidimensionality of a test—the extent to which items load on a single common factor. This section is devoted to showing why α is not useful as an index of test homogeneity.

Homogeneity and Internal Consistency

I follow the suggestion of Green, Lissitz, and Mulaik (1977) in using "homogeneity" to mean *unidimensionality* because this seems to be the term's usual meaning, but there is some ambiguity in the literature (see McDonald, 1981, for discussion of this). The term *internal consistency* is also ambiguous, but I will use it, again following Green et al., to refer to interrelatedness of items—a high positive mean interitem correlation with few or no negative intercorrelations would indicate that a test is internally consistent. Generally speaking, a test that is homogeneous will be internally consistent, but a test that is internally consistent is often not homogeneous. α does not make a good index of either of these constructs for several reasons.

Homogeneity and Test Length

Cronbach (1951) emphasized that α should not be used as an index of test homogeneity,

> Conceptually, it seems as if the 'homogeneity' or 'internal consistency' of a test should be independent of its length. A gallon of homogenized milk is no more homogeneous than a quart. α increases as the test is lengthened. (p. 323)

He went on to derive an index unrelated to test length. His admonition was not taken to heart by all who followed. Green et al. (1977) were able to present many instances in which other writers mistook α for a homogeneity index, and McDonald (1981) and Cortina (1993) also discussed the problem at some length. The problem of test length is familiar to most readers because

of its use in stepping up split-half reliability estimates. We can conclude, as did Cronbach, that if new items added to a test are positively correlated with items already in the test, test reliability will increase and α will increase. In our terminology, test reliability increases with test length when a test is internally consistent, whether the test is homogeneous or not.

A Demonstration of the Failure of α as an Index of Homogeneity

We will now consider a model for a test that violates the assumptions of essential tau-equivalence and homogeneity. Figure 3 contains a path diagram indicating the items of a test. The test score, X, was left off the path diagram to make it easier to read. If it had been included, there would be an additional arrow with a path coefficient of 1.0 from each item score to the X variable. It can be seen from Figure 3 that a score on this hypothetical test is a sum of m uncorrelated subtests, each having k essentially tau-equivalent items. That is, the test has m orthogonal (uncorrelated) true-score factors with k items loading on each factor, a total of $m \cdot k$ items. Each item loads on only one factor, and all items are assumed to have the same true-score variance and the same error variance. Interitem correlations are zero between items of different subtests. For items of the same subtest, the reader should confirm that the correlation is equal to the item reliability, for example, $\mathrm{Var}(t)/\mathrm{Var}(x_{ij}) = r$. It is then possible to write an equation for α in terms of m, k, and r.

$$\alpha = \frac{m \cdot k}{(m \cdot k - 1)} \cdot \frac{r \cdot (k-1)}{(1 + r \cdot [k-1])} \tag{9}$$

Cortina (1993) did not present this equation, but he seems to have used it to produce the values in his Table 2. The equation was presented, though in a slightly different form, by Wherry and Gaylord (1943).

Heuristic value of Equation 9. The hypothetical set of test items shown in Figure 3 was selected because it shows such a clear departure from homogeneity. No reasonable homogeneity index should show this test to be homogeneous. On the other hand, the test does show some internal consistency—all interitem correlations are positive or zero. Equation 9 allows us to show the quantitative relation of α to various test parameters such as test dimensionality, number of items, and interitem correlation, within the simple framework of the model in Figure 3. First, when r is zero, α is zero, as it should be. In our model, when all interitem correlations are zero, true-score variance also goes to zero, and both α and test reliability should go to zero. Second, α increases with r and reaches a maximum of $(m \cdot k - m)/(m \cdot k - 1)$

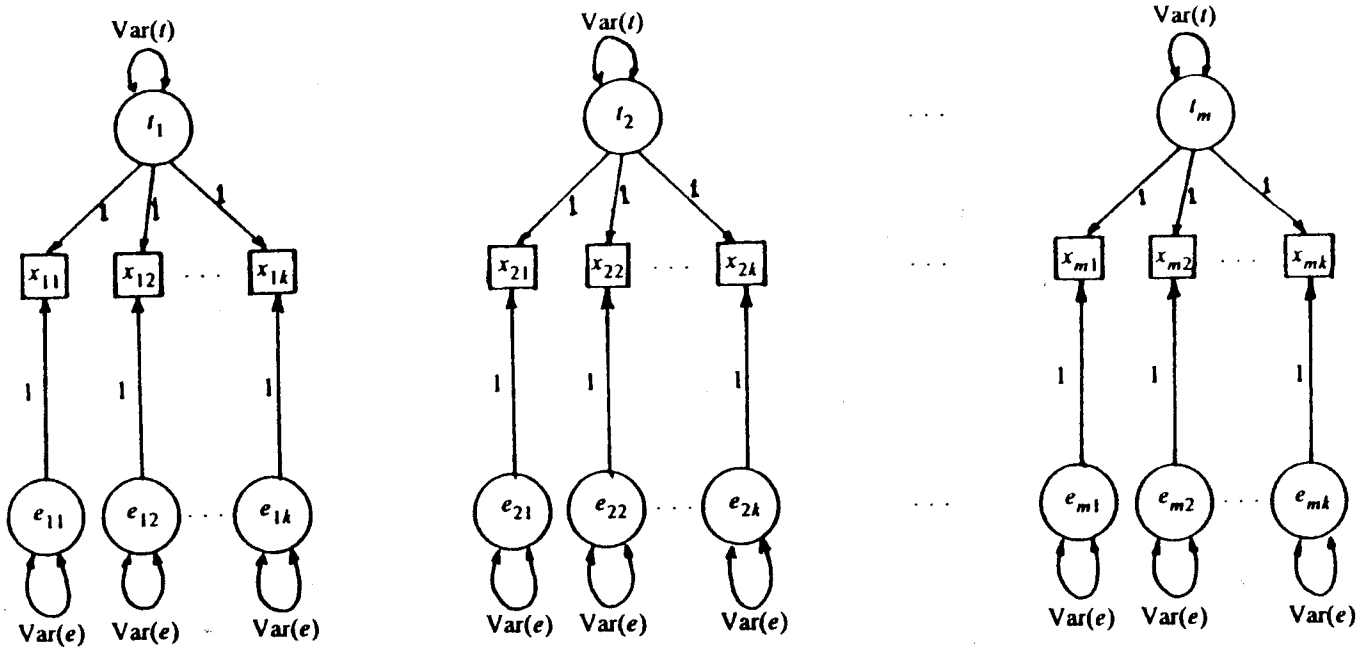

FIGURE 3 Path diagram showing the relations of items from a heterogeneous test with m orthogonal true-score factors.

when r reaches 1.0. A large value of r should make the orthogonal factors more and more obvious and a good index of homogeneity should detect them more readily. α fails miserably in this regard. For example, with 5 true-score factors and 10 items per factor, in the extreme case where r equals 1.0 and the factors are perfectly distinguishable in any sample data, α will equal .92. Of course, internal consistency and reliability both increase with r, thus making sense of the increase in α. Third, in our model, with any fixed number of factors, α is highly dependent on the number of test items. For example, a test with 4 orthogonal factors and items loading on the same factor intercorrelated .20, α would be .66 with 10 items per factor, .80 with 20 items per factor, and .91 with 50 items per factor. This is a highly undesirable property of any homogeneity index because the "evidence" favoring the false hypothesis of homogeneity increases with information (number of test items). Of course, we can understand the relation of α to number of items if we interpret α as an index of reliability rather than homogeneity. Finally, we can see that α is amazingly insensitive to the number of orthogonal factors. For example, if there are 10 items per factor and r is held constant, as the number of factors is increased from 1 to 5, α will decrease by only 8% of its value, for example, from .75 to .69. The number of items in this last example increased from 10 to 50, but test reliability actually remained constant (the reader may wish to verify this) because the added items were uncorrelated with those already present. Thus, the failure of α to decrease more steeply with decreasing homogeneity can not be accounted for by increasing test reliability. Again, α is not a homogeneity index.

Criticism of Equation 9. Some readers might argue that Equation 9 tells us little about α because the assumptions of the model from which it is derived are absurdly restrictive. This argument is unconvincing. If an index of homogeneity can't perform appropriately under these simple conditions, when can it? The restrictions of orthogonality, equal loadings, equal error variances, and so forth, should only improve the performance of a homogeneity index. Thus, the conclusion that α has few of the properties of a homogeneity index, is conservative.

CONCLUSION

Coefficient alpha, computed on real test data, yields an estimate of a lower bound on test reliability. α is a very helpful index in test construction for tests that are designed to measure a single trait dimension. However, α has little or no value as an index of test homogeneity or unidimensionality, and it can badly underestimate reliability when the test is not unidimensional.

Researchers often find that they can't retain all their test items when fitting a structural equation model because of the large number of parameters and because of highly nonnormally distributed item scores. Summing item scores can reduce the number of observed variables dramatically, yielding a manageable model with roughly normally-distributed variables. After reducing the number of variables, the researcher will often need an estimate of reliability that can be used to make the measurement model identified. This can be a risky business because an incorrect reliability estimate will lead to biased estimates of other model parameters.

The simplest example of the use of reliability estimates in structural equation models is in the correction for attenuation of correlation due to measurement error (see Bollen, 1989; Nunnally, 1978). The observed correlation is divided by the geometric mean (product of square roots) of the reliability coefficients of the two variables. The result is an estimate of the correlation that would be found in the absence of measurement error. If the reliability estimates are too low, as they are expected to be when α_{est} is used as the estimator, the method will overcorrect and the researcher will be led to believe that the underlying relation is stronger than it truly is.

This problem of bias arises for much the same reason in more sophisticated structural models. For example, if αs are used to estimate the relative contributions of ξ and δ to the variances of X variables, the path coefficients (elements of Γ) from ξ to η will be overestimated. This could cause elements of Γ to differ significantly from zero, thus inflating the Type 1 error rate. It is important to bear this in mind and interpret results with appropriate caution when α has been used to estimate reliability.

ACKNOWLEDGMENTS

This work was supported by National Institute of Mental Health postdoctoral training grant MH55030.

I thank James Wollack, Loren Chapman, and Timothy Gallagher for their helpful comments on an earlier draft of this article, and Ed Rigdon, Pamela Madden, and Karl Jöreskog for their suggestions regarding the LISREL script.

REFERENCES

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12*, 1–16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99–105.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93–103.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827–838.

Jöreskog, K. G., & Sörbom, D. (1994). *LISREL 8 user's reference guide*. Chicago, IL: Scientific Software International.

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

McArdle, J. J., & Boker, S. M. (1990). *RAMpath path diagram software*. Denver, CO: Data Transforms.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.

Neale, M. C. (1994). *Mx: Statistical modeling. User's Guide*. Department of Human Genetics, Box 3, Medical College of Virginia, Richmond, VA.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. Dordrecht, The Netherlands: Kluwer Academic.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13.

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

Pandey, T. N., & Hubert, L. (1975). An empirical comparison of several interval estimation procedures for coefficient alpha. *Psychometrika, 40*, 169–181.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split halves. *Harvard Educational Review, 9*, 99–103.

Wherry, R. J., & Gaylord, R. H. (1943). The concept of test and item reliability in relation to factor pattern. *Psychometrika, 8*, 247–26

APPENDIX

The LISREL script shown here will compute α for a *k*-item test from the variance–covariance matrix of the *k* items. It will create a file called ALPHA.OUT that contains, in order, the estimated true-score variance, the test variance, and coefficient alpha. If you are using LISREL 7 or an earlier version, remove the CO (constraint) lines, delete NE=3 and PS=ZE from the MO line, and delete GA=ALPHA.OUT from the OU line. α can then be computed as $k^2 \cdot \text{PHI}(1) / [k^2 \cdot \text{PHI}(1) + k \cdot \text{TD}(1)]$. The remaining LISREL output, fit statistics, chi squares, and so forth, should be ignored.

```
LISREL SCRIPT FOR COMPUTING COEFFICIENT ALPHA
; REPLACE EVERY K IN THIS SCRIPT (EXCEPT IN "NK=1")
; WITH THE NUMBER OF ITEMS
DA NI=K NO=1000
CM
;
; ENTER COVARIANCE MATRIX HERE OR USE FI=FILENAME
;
MO NK=1 NX=K NE=3 PS=ZE
VA 1.0 LX(1)-LX(K)
EQ TD(1)-TD(K)
CO GA(1)=K*K*PH(1)
CO GA(2)=K*K*PH(1)+K*TD(1)
CO GA(3)=GA(1)*GA(2)**-1
OU ME=UL GA=ALPHA.OUT
```

The Mx script that follows will compute coefficient alpha from a covariance matrix and will place it on the second line of a file named alpha.out. Other output of the program, fit statistics, and so forth, should be ignored.

```
Mx script to compute coefficient alpha for a k item test
! replace every occurrence of the letter k with the number of items
DAta CAlculate NGroups=1
MATrices
S SYmm k k
J UNit k 1
O UNit 1 1
m SYmm 1 1
COmpute (m%(m-O))@(O-\tr(S)%(J'*S*J)) /
MAtrix S !add FI=filename here for file with k x k covariance matrix
!
! or type covariance matrix here if it isn't in a file
!
MAtrix m
k
OU MX%E=alpha.out
```