SA Journal of Industrial Psychology, 2006, 32 (4), 34-40
SA Tydskrif vir Bedryfsielkunde, 2006, 32 (4), 34-40

COEFFICIENT ALPHA: UNNECESSARILY AMBIGUOUS; UNDULY UBIQUITOUS

G.K. HUYSAMEN
gerthuysamen@mweb.co.za
Gordon Institute of Business Science,
University of Pretoria,
and
Department of Psychology,
University of the Free State,

ABSTRACT

By means of simple numerical examples it is demonstrated how the value of coefficient alpha is determined by the size and sign of the item intercorrelations, the dimensionality of the items, and the number of items. Relatively high alphas may be obtained for multidimensional item data (as obtained under a violation of essential tau-equivalence). Dimensionality analyses should therefore precede internal-consistency analyses if unidimensional tests are required. Such analyses also may alert one to the possibility of negative alpha values. Alternative coefficients are mentioned for situations when essential tau-equivalence is in doubt.

Key words
Coefficient alpha, internal consistency

When Cronbach (1951) introduced coefficient alpha more than half a century ago, he correctly anticipated the usefulness of this index of internal consistency. It has proven to be the most popular reliability estimation method by far (Hogan, Benjamin & Brezinzki, 2000), and social sciences citations of Cronbach's 1951 article run into several hundreds per year (Cronbach, 2004). Cortina (1993) refers to this coefficient's use not only in the various areas of psychology, but also in sociology, statistics, medicine, counseling, nursing, economics, political science, criminology, gerontology, broadcasting, anthropology and accounting. Undoubtedly this coefficient's popularity may be attributed to the fact that it doesn't require more than one test administration (as does the test-retest method), or more than one parallel test form (as does the parallel-forms method), or the splitting of a test into two parallel halves (as do the split-half methods).

Not only is Cronbach's 1951 article the source of the highly popular coefficient alpha, but it still continues to stimulate all kinds of methodological comments and developments. There has been a continuous stream of attempts to derive alternative coefficients for situations where the assumptions underlying coefficient alpha have been relaxed (cf. Lucke, 2005). Other methodological contributions (e.g., those of Cortina, 1993; Henson, 2001; Schmitt, 1996; Streiner, 2003 and Thompson, 2003) have had a more modest aim, namely, to explain some of the anomalies and misconceptions associated with coefficient alpha. These contributions addressed, amongst others, the possibility of relatively high alpha values for multidimensional item data and the occurrence of negative alpha values. However, it would appear that these messages still haven't reached all of those who have access to computer programmes for the computation of coefficient alpha – it is not uncommon to find applications of coefficient alpha as a test of unidimensionality or as an internal-consistency index across subtests (of relatively independent constructs) in some locally published research.

To explain the intricacies of coefficient alpha, the authors referred to above used numerical examples of item variance-covariance (VCV) matrices which may be less than helpful in explaining how such aberrant matrices have come about in the first place. The purpose of the present article is to revisit these earlier contributions but to start off the explanations involved in terms of simple item data matrices rather than item VCV matrices. In the process, some inconsistencies and omissions in at least some of these contributions will be noted and clarified. A clear understanding of coefficient alpha may prevent its incorrect use and interpretation. Finally, some of the alternatives to coefficient alpha for situations in which some of its assumptions do not hold will be mentioned. But first, a review of the assumptions, derivation and computation of this coefficient is presented.

ASSUMPTIONS AND DERIVATION OF COEFFICIENT ALPHA

The computation of coefficient alpha is indicated when scores on a composite test (the total scores on a set of items) are to be interpreted as indicative of some construct that presumably underlies all the items. In other words, if one person obtains a higher total score than another on such a composite test, the test user would like to conclude that the former person occupies a higher position than the latter on whatever construct the test (questionnaire, etc.) is designed to measure. For this reason, it makes no sense to compute coefficient alpha on the sum of individuals' responses to, for example, a set of biographical questions (age, marital status, salary, etc.) even if the responses to the various items have been converted into some common metric. To interpret the total scores on such a set of items would make no sense as they cannot be considered to be reflecting a common underlying construct (e.g., something like a "biographical trait"). In structural equation modelling terminology, coefficient alpha may be said to become relevant when we are dealing with effect indicators rather than causal indicators of constructs (Bollen & Lennox, 1991).

Classical test theory decomposes an individual's observed test score, *X*, into a true-score and an error-score component. The true score, *T*, is defined as a person's mean over an infinitely large number of independent applications of the test. A person's error score, *E*, on a particular test application is the deviation of the observed score for that application from his or her (constant) true score. Under certain assumptions (e.g., that error scores are uncorrelated), it can be shown that the observed-score variance (over persons) is equal to the sum of the true-score variance and the error-score variance:

$$\sigma^2_X = \sigma^2_T + \sigma^2_E$$

Reliability, then, is defined statistically as the ratio of true-score variance to observed-score variance:

$$\rho_{rel} = \sigma^2_T/\sigma^2_X, \quad (1)$$

Equivalently, it may be defined as the correlation between strictly parallel tests,

$\rho_{rel} = \rho_{jk}$,

where strictly parallel tests *j* and k are defined to have the same true score for any particular person *i*: $T_{ij} = T_{ik}$. Such tests have equal observed-score variances, equal true-score variances, and equal error-score variances. It can be shown (Lord & Novick, 1968) that for any pair of parallel tests, *j* and *k*, with true scores *t*′ and *t*″, respectively, the true-score variances and true-score covariance are equal:

$$\sigma^2_{t'} = \sigma^2_{t''} = \sigma^2_{t' t''}. \quad (2)$$

Moreover, it can be shown that the true-score covariance of any pair of parallel tests is equal to the observed-score covariance of such tests:

$$\sigma_{t't''} = \sigma_{jk}. \quad (3)$$

The derivation of coefficient alpha assumes that the components (items or subtests) comprising the total test are essentially tau-equivalent, which is a less restrictive assumption than parallelism: individuals' true scores on essentially tau-equivalent components are allowed to differ by a constant, *c*: $T_{ij} = c + T_{ik}$, so that true scores on essentially tau-equivalent components, although unequal, are perfectly correlated. If, for example, c = 0,68, one person may have true scores of 16,23 and 16,91 on such components whereas another may have true scores of 12,19 and 12,87, respectively, on them. (So, such components may be said to be essentially true-score equivalent, hence the term essentially tau-equivalent, as true scores often are denoted by τ, pronounced *tau*.) Whereas parallelism implies equally difficult items or tests, essential tau-equivalence allows for differences in difficulty. For example, a spelling test containing words such as *computer* and *complication* may show higher true scores (be less difficult) for individuals than one comprising of words such as *commission* and *accommodation*, and yet these tests may be highly correlated. As the addition of a constant (such as *c*) leaves variances and covariances unaffected, Equations (2) and (3) apply with equal force to essentially tau-equivalent components. (However, the observed-score variances of essentially tau-equivalent components, unlike those of parallel components, may differ.)

To understand coefficient alpha as a function of the variances and covariances of the items comprising a test, it is instructive to consider the VCV matrix for such items. For a test of *J* items, there are *J* item variances and $J(J-1)$ covariances (cf. Tables 1b, 2b, 3b, and 4b for sample data). The diagonal elements (indicated in bold) of such a $J \times J$ matrix are the variances of the *J* items, and the $J(J-1)$ off-diagonal elements are the covariances. The CVC matrix is symmetrical, by which is meant that for every pair of items (e.g., the 2nd and 5th items) there are two (equal) covariance terms, one (namely, s_{25}) situated in the upper right triangle (in the cell formed by the second row and the fifth column) and its identical twin member (s_{52}) located in the lower left triangle (in the fifth row, second column). Because of this equivalence the covariance terms in only one of these triangles are typically shown. In common with the variance of any linear combination of J variables, the variance of total test scores is equal to the sum of all the entries in the VCV matrix for the items comprising the test. So, the variance of the total observed score *X*, where $X = x_1 + x_2 .... x_J$, may be given as:

$$\sigma_X^2 = \Sigma\sigma_j^2 + 2\Sigma_j\Sigma_k\sigma_{jk},\ k = j + 1, \quad (4)$$

where $\Sigma\sigma_j^2$ = the sum of the *J* item variances; σ_{jk} = is the covariance of items *j* and *k*; and the double summation sign, $\Sigma_j\Sigma_k$, in this and in all subsequent applications, in view of $k = j + 1$, means that all the $J(J-1)/2$ entries in only the upper right triangle of the VCV matrix are summed. (Multiplication by 2 takes care of the equivalent covariances in the lower left triangle.)

In the derivation of coefficient alpha (cf. Novick & Lewis, 1976) as the reliability of a composite test, Equation (4) may be used for the observed-score variance in Equation (1). Next, an equation for σ^2_T, the variance of total true scores, *T* (where $T = t_1 + t_2 + ... + t_J$), in terms of an observable quantity, is required. Equations (2) and (3) make it possible to express σ^2_T in terms of the sum of item observed-score covariances,

$$\sigma^2_T = [J/(J-1)][2\Sigma_j\Sigma_k\sigma_{jk}],$$

which, upon substitution into Equation (1), yields:

$$\alpha = [J/(J-1)][2\Sigma_j\Sigma_k\sigma_{jk}/\sigma^2_X], \text{ where } k = j + 1. \quad (5)$$

If one solves Equation (4) for $2\Sigma_j\Sigma_k\sigma_{jk}$ and substitutes the result into Equation (5), the more familiar version of coefficient alpha is obtained:

$$\alpha = [J/(J-1)][1 - (\Sigma\sigma_j^2/\sigma^2_X)],$$

which may be easier for computational purposes. However, Equation (5) proves to be more helpful in showing how coefficient alpha is determined by the size and signs of the item covariances.

Computation and interpretation of coefficient alpha

To facilitate the computations in the following numerical examples, dichotomous items will be used. Coefficient alpha is applicable to both dichotomous and multi-point items (such as those in typical Likert scales), whereas the algebraically equivalent Kuder-Richardson formula 20 is applicable to dichotomous scores only (Cronbach, 1951). The variance of a dichotomous item *j*, written in Roman symbols now to denote sample statistics, is equal to

$$s^2_j = p_j(1 - p_j), \quad (6)$$

where p_j is the proportion of participants who answered item *j* correctly (or endorsed it, in typical-performance measurement). In all the following numerical examples every item is passed (or endorsed) by half of the participants. As a result, these items show the maximum variance (for dichotomous variables but not for items with more than two categories) of 0,25 so that the effect of item variance is kept constant. The covariance of items *j* and k is equal to

$$s_{jk} = p_{jk} - p_j p_k, \quad (7)$$

where p_{jk} is the proportion of participants who passed (or endorsed) both items *j* and *k*.

Consider the example of the item data matrix in Table 1a, where rows refer to participants (A, B, etc.) and columns to items (1, 2, etc.). In this example there is perfect consistency in responding among the performances of the eight individuals on the five items. Any person (such as A) who passes Item 1, is sure to pass all of the other four items; any person (such as E) who fails Item 1, is bound to fail all of the remaining four items. The VCV matrix in Table 1b indicates that in terms of Equation (6), every item's variance is equal to 0,50(0,50) = 0,25. In terms of Equation (7) the covariance of every pair of items, say that of Items 1 and 2, is equal to 0,50 – 0,50(0,50) = 0,25, p_{12} being equal to 0,50 because four (A, B, C, and D) of the eight participants passed both items 1 and 2. (Notice that a covariance of 0,25 corresponds to a perfect item intercorrelation as $r_{12} = s_{12}/s_1s_2 = 0,25/(\sqrt{0,25})(\sqrt{0,25}) = 1,00$.) As a result, the VCV matrix contains five variance terms (there being five items), each equal to 0,25, so that $\Sigma s^2_j = 5(0,25) = 1,25$. There are $J(J-1) = 5(4) = 20$

covariance terms (recall that those in the lower left triangle are not shown), each equal to 0,25, so that $2\Sigma_j\Sigma_k s_{jk} = 20(0,25) = 5$. As a result, in the sample analog of Equation (4) in which population parameters σ^2_j and σ_{jk} are replaced by sample statistics s_j^2 and s_{jk}, respectively, the total-test variance is equal to 1,25 + 5 = 6,26. By Equation (5), the sample estimate of coefficient alpha for the five items is equal to = [5/4][5/6,25] = 1,00. A value of 1,00 for coefficient alpha is quite apt as the item matrix on which it is based reflects perfect consistency in responding.

TABLE 1
AN EXAMPLE OF COMPLETE INTERNAL CONSISTENCY

1.a: Item matrix

	1	2	3	4	5
A	1	1	1	1	1
B	1	1	1	1	1
C	1	1	1	1	1
D	1	1	1	1	1
E	0	0	0	0	0
F	0	0	0	0	0
G	0	0	0	0	0
H	0	0	0	0	0

1.b: Variance-covariance matrix

	1	2	3	4	5
1	**0,25**	0,25	0,25	0,25	0,25
2		**0,25**	0,25	0,25	0,25
3			**0,25**	0,25	0,25
4				**0,25**	0,25
5					**0,25**

Table 2a presents the other extreme as there is no consistency among the individuals' scores on the items in this example. If a person passes any particular item, there is no way of telling how he or she would fare on any other item. For example, whereas both individuals A and B pass Item 1 and also pass Item 2, individuals C and D, who also pass Item 1, fail Item 2, and so on. Now, if we follow the same procedures as before, the covariance of every pair of items is found to be equal to 0,25 – 0,50(0,50) = 0,00. The VCV matrix consists of five variance terms, each of which is equal to 0,25, and 20 covariance terms, each equal to 0. The variance of the total test, therefore, is equal to 5(0,25) + 20(0,00) = 1,25, and coefficient alpha equals [5/4][0/1,25] = 0. Once again, the obtained value is fitting because the value of zero reflects complete inconsistency in responding.

TABLE 2
AN EXAMPLE OF COMPLETE INTERNAL INCONSISTENCY

1.a: Item matrix

	1	2	3	4	5
A	1	1	1	1	0
B	1	1	0	1	1
C	1	0	1	0	0
D	1	0	0	0	1
E	0	1	1	0	1
F	0	1	0	0	0
G	0	0	1	1	1
H	0	0	0	1	0

1.b: Variance-covariance matrix

	1	2	3	4	5
1	**0,25**	0,0	0,0	0,0	0,0
2		**0,25**	0,0	0,0	0,0
3			**0,25**	0,0	0,0
4				**0,25**	0,0
5					**0,25**

An intuitively meaningful definition of coefficient alpha is that it is the mean of all possible split-half reliability coefficients that could be determined for a test (Cronbach, 1951). The value obtained for any split depends to some extent on that particular split, for example, whether it consists of the first half of the items against the second half, or all odd-numbered items against all even-numbered items. As the mean of all such split-half reliabilities, coefficient alpha is a more stable internal-consistency estimate than that based on any particular split. However, the present split-half reliabilities are not those obtained by correlating two parallel split halves and stepping up the obtained value by means of the Spearman-Brown formula. Instead, it is the mean of the Rulon-Flanagan split-half reliabilities that are computed by:

$$r_{rel} = (4s_{12})/s_X^2,$$

where s_{12} is the covariance of two essentially tau-equivalent test halves (Feldt & Brennan, 1989).

Finally, if we bear in mind the relation between correlation and covariance, $r_{12} = s_{12}/s_1s_2$, it follows that coefficient alpha somehow has to be related to the item intercorrelations. If we define the average item intercorrelation, $\bar{r}_{jk}$, as follows,

$$\bar{r}_{jk} = [2\Sigma_j\Sigma_k s_{jk}/J(J-1)]\ /[\Sigma_j s_j^2/J],$$

coefficient alpha of the composite of J items can be shown to result from applying the Spearman-Brown formula to $\bar{r}_{jk}$ (Feldt & Brennan, 1989):

$$\alpha = J\,\bar{r}_{jk}/[1 + (J-1)\,\bar{r}_{jk}]. \qquad (8)$$

COEFFICIENT ALPHA AND DIMENSIONALITY

In factor-analytic terms, the observed-score variance of a composite test may be broken down into common-factor variance (where common factors are those on which several items show high loadings), specific-factor variance (specific factors being those on which only individual items load non-negligibly), and residual error variance. Coefficient alpha is related to the proportion of common-factor variance (as opposed to specific-factor and residual error variance) but for a comprehensive explication of this relationship, the reader is referred to Cronbach (1951) and McDonald (1981). Cronbach (1951) pointed out that among the common factors the first one is likely to be a general factor even though all items may not uniformly show high, or even nonzero, loadings on it. As the number of such items increases, the proportion of the total test-score variance due to such a general factor and, hence, coefficient alpha, increases. This has led to alpha being described as a measure of first-factor saturation (Cortina, 1993), in other words, the extent to which most, if not all, items load appreciably on the first factor.

However, Green, Lissitz and Mulaik (1997) showed that the presence of such a general factor is not a prerequisite for high alpha values. They artificially constructed an example of a ten-item, five-factor test in which (i) each (factorially complex) item had a loading of 0,67 on each of two orthogonal factors, (ii) no pair of items showed such loadings on the same two factors, and (iii) each item's loadings on the remaining three factors were zero. (For example, each of Items 1 to 4 showed a loading of 0,67 on Factor I and numerically the same loading on each of Factors II to V, respectively.) Thirty of the 45 unique item pairs showed correlations of 0,45 and each of the remaining 15 item pairs had a correlation of zero, signifying a violation of essential tau-equivalence. The mean item intercorrelation over all 45 item pairs was 0,30 and, by Equation (8), coefficient alpha for the composite of ten factorially complex items was equal to 0,81. As the commonality of each item was equal to $0,67^2 + 0,67^2 = 0,90$, common-factor variance was bound to occupy the major share of observed-score variance and so resulted in a high coefficient alpha.

Cronbach (1951) emphasised that for a test to be interpretable, all items need not be factorially similar but instead that the first common factor should account for a large proportion of test variance. Obviously, a test such as Green et al.'s (1977) example with five equally strong factors may be rather difficult to interpret. As a matter of fact, Hattie (1985) considers unidimensionality to be a prerequisite for interpretability. (For the analyses of multidimensional item data, including factorially complex item data, the interested reader is referred to McDonald, 1999, and Lucke, 2005.) And if unidimensionality is sought, the preceding paragraph clearly suggests that coefficient alpha should not be considered for gauging unidimensionality. Rather, it is generally recommended that a dimensionality analysis should precede the calculation of coefficient alpha (Gerbing & Anderson, 1988). Hattie (1985) discusses among others the use of factor and component analysis and latent trait analyses for this purpose and Schmitt (1996) refers to the increasing use of structural equation modelling in this regard.

DETERMINANTS OF THE SIZE OF COEFFICIENT ALPHA

The numerator of the key ratio, $2\Sigma_j\Sigma_k\sigma_{jk}/\sigma^2_X$, in Equation (5) for coefficient alpha consists of the sum of the covariances whereas, in terms of Equation (4), the denominator consists of the sum of the covariances plus the sum of the item variances. In other words, for a fixed sum of item variances, the size of coefficient alpha will increase as the sum of the covariances increases. Although the variance of an item places a limit on its covariance with any other item, as the number of items increases, the number of covariance terms increases at a faster rate (than does the number of variances), and so allows the present ratio (and, hence, coefficient alpha) to increase as well. The VCV matrix for a five-item test ($J = 5$), has only five variance terms but $J(J-1) = 5(5-1) = 20$ (i.e., four times more) covariance terms; for a six-item test there are six variances but 30 (i.e., five times more) covariances. But more important than the number of items and the attendant number of covariances is the presence of covariances that are both positive and large in sign.

More specifically, Equation (8) shows that for a particular average item intercorrelation there is a curvilinear relationship between test length and reliability. As a test is lengthened by the addition of parallel or essentially tau-equivalent components, the reliability of the resulting composite increases but sooner or later it reaches a point where such increase tapers off. The lower the reliability of the initial test, the greater the number of times it should be lengthened before this point is reached. For example, if the initial test has a reliability of 0,80, lengthening it six times would yield a reliability of 0,96 but any further increases in test length wouldn't raise reliability appreciably. However, if the initial reliability is 0,10 only, lengthening it 12 times would raise its reliability to only 0,57, and the point of inflection then still hasn't been reached.

Cortina (1993) demonstrated the manner in which coefficient alpha is affected by the number of items, the average item intercorrelation and the number of dimensions present among the items. He calculated alphas for 6-, 12- and 18-item tests for each of three situations, namely, where the average item intercorrelation was 0,30, 0,50 and 0,70, respectively. For an average item intercorrelation of 0,30, coefficient alpha for these three test lengths is equal to 0,72, 0,84 and 0,88, respectively. For an average item intercorrelation of 0,70, the corresponding increases in the number of items result in coefficients of 0,93, 0,96 and 0,98, respectively. An increase in the items benefits alpha also in the case of multidimensional data. For example, Cortina's data indicate that in the case of two orthogonal factors, alpha for a six-item test for the three situations (in which the factorially simple items within a particular cluster now show the mean intercorrelations mentioned), are 0,45, 0,60 and 0,70, respectively. Under the same conditions, 18 items raised the coefficients to 0,75, 0,85 and 0,90, respectively.

The data in Table 3 demonstrate how alpha is affected when factorially simple items divide into two clusters that each measures a distinct construct. Items 1 through 5 behave much like any item in Table 1 and so do Items 6 through 10. If a person passes Item 1 he or she is likely to pass Items 2 to 5 as well; if he or she fails Item 1, he or she is likely to do the same on Items 2 to 5 as well. Similarly, a person's performance on Item 6 predicts pretty well his or her performance on Items 7 to 10. However, there is no similar consistency between individuals' performances on the first cluster (Items 1 to 5) and their performances on the second cluster (Items 6 to 10). Some individuals (such as A) who pass Items 1 to 5 also pass Items 6 to 10 but others (such as B) who pass the first five items, fail the last five items. This lack of consistency shows up in the uniformly poor covariances (printed in italics in Table 3b) between any item from the first cluster and any item from the last cluster. For example, the covariance of Items 1 and 10 is equal to 0,30 – 0,50(0,50) = 0,05, which translates into an item intercorrelation of $0{,}05/(\sqrt{0{,}25})(\sqrt{0{,}25}) = 0{,}20$. There are 5 × 5 = 25 such item pairs that clearly fail to meet the assumption of essentially tau-equivalence. Each cluster of five items shows a coefficient alpha of 0,88 (on its own). However, if one treats the ten items as a single composite, the within-cluster covariances sum to 6,00 whereas the between-cluster covariances sum to only 0,10 and so restrict the total sum of covariances to 6,10. Total test variance is equal to 8,60, and coefficient alpha amounts to (10/9)(6,10/8,60) = 0,79. This is a quite respectable value, that is, despite the fact that there are two clusters of items that obviously represent two different dimensions. This comes about because the within-cluster covariances (in regular print) have such high values relative to the between-cluster covariances that the mean over all $J(J-1)$ covariance terms is still high and coefficient alpha for the composite test has an appreciably high value.

TABLE 3
AN EXAMPLE OF HETEROGENOUS ITEM DATA WITH HIGH COEFFICIENT ALPHA

3.a: Item matrix

	1	2	3	4	5	6	7	8	9	10
A	1	1	1	1	1	1	1	1	1	1
B	1	1	1	1	1	0	0	0	0	0
C	1	1	1	1	1	1	1	1	1	1
D	1	1	1	1	1	0	0	0	0	1
E	0	0	0	0	0	1	1	1	1	1
F	1	0	0	0	0	0	0	0	1	0
G	0	1	0	0	0	1	1	1	1	1
H	0	0	1	0	0	0	0	1	0	0
I	0	0	0	1	0	0	1	0	0	0
J	0	0	0	0	1	1	0	0	0	0

3.b: Variance-covariance matrix

	1	2	3	4	5	6	7	8	9	10
1	**0,25**	0,15	0,15	0,15	0,15	*-0,05*	*-0,05*	*-0,05*	*0,05*	*0,05*
2		**0,25**	0,15	0,15	0,15	*0,05*	*0,05*	*0,05*	*0,05*	*0,05*
3			**0,25**	0,15	0,15	*-0,05*	*-0,05*	*0,05*	*-0,05*	*0,05*
4				**0,25**	0,15	*-0,05*	*0,05*	*-0,05*	*-0,05*	*0,05*
5					**0,25**	*0,05*	*-0,05*	*-0,05*	*-0,05*	*0,05*
6						**0,25**	0,15	0,15	0,15	0,15
7							**0,25**	0,15	0,15	0,15
8								**0,25**	0,15	0,15
9									**0,25**	0,15
10										**0,25**

EVALUATING COEFFICIENT ALPHA'S SIZE

From the preceding section it follows that a blanket, rule-of-thumb statement to the effect that coefficient alpha should have a particular value may be misleading. Cortina's (1993) data show that a (relatively high) coefficient of 0,85 may be obtained in three quite different situations: (i) 18 items comprising of two clusters of nine items each, each measuring a different construct, and with an average item intercorrelation of 0,50 within each cluster; (ii) 18 items measuring a single construct, with an average item intercorrelation of only 0,30; or (iii) only six items measuring a single construct, with an average item intercorrelation of 0,50.

Schmitt (1996, p. 350) uses the attenuation formula and argues that "(e)ven with a reliability as low as .49, the upper limit of validity is 0,70". The attenuation formula states that validity as the correlation between a test and a criterion is equal to (i) the product of the correlation between the true scores on the test and those on the criterion, (ii) the square root of the reliability of the test, and (iii) the square root of the reliability of the criterion:

$$\rho_{XY} = \rho_{XY}^2 \sqrt{\rho_{XX'}} \sqrt{\rho_{YY'}}, \quad (9)$$

where X represents the true scores on the test, X; Y indicates the true scores on the criterion, Y; and where $\rho_{XX'}$ and $\rho_{YY'}$ denote the reliability of the test and of the criterion, respectively. However, Schmitt fails to mention that for Equation (9) to show that a reliability of 0,49 is sufficient for an upper limit, ρ_{XY} = 0,70, one has to assume that the correlation between the true scores on the two variables equals unity (ρ_{XY} = 1,00), and that the criterion is perfectly reliable ($\rho_{YY'}$ = 1,00). Both of these assumptions are unrealistic in the extreme. If one substitutes ρ_{XY} = 0,80, and $\rho_{YY'}$ = 0,90 in Equation (9), it turns out that $\rho_{XX'}$ has to be equal to 0,85 for ρ_{XY} to be equal to 0,70. This result seems to cast serious doubt on the acceptability of rules of thumb that tolerate alpha values as low as 0,70.

Several authors point out that a high alpha may be obtained simply by focusing on a narrowly defined area which doesn't correlate with any criterion of importance. Streiner (2003) cautions that higher values may be obtained through an unnecessary duplication of content across items, for example, by phrasing the same question in many different ways. Although such redundancy certainly may result in a high alpha, a high alpha does not necessarily reflect such redundancy.

There is the danger that thumb-sucking exercises regarding the appropriate size for coefficient alpha may conceal a more important issue, namely, that coefficient alpha is registering only specific-factor or content-specific measurement error (apart from random response error). To the extent that measurement error due to instability in responding over time (transient error) is present, coefficient alpha presents an overestimate of reliability. Textbooks diligently spell out that if the assumption of essentially tau-equivalence does not hold, coefficient alpha represents a lower bound to reliability. It is less often mentioned that if transient error or correlated error scores are present (and it is a good bet that generally such flaws do prevail), coefficient alpha overestimates reliability. A more inclusive view would be that whether coefficient alpha is an underestimate or an overestimate depends on which one of these assumptions is more seriously violated (Becker, 2000; Komaroff, 1997). So, in the presence of transient error, obtaining an alpha even as high as one of 0,90 may be regarded as a Pyrrhic victory.

THE OCCURRENCE OF NEGATIVE VALUES FOR COEFFICENT ALPHA

More than half a century ago, Cronbach and Hartmann (1954, p. 344) pointed out that "negative internal-consistency coefficients may be expected when negatively correlated factors are thrown together in one test". More specifically, a negative coefficient alpha is obtained when the (absolute value of the) sum of the negative covariances exceeds the sum of the positive covariances. For this to occur, the items should divide into (at least) two clusters of one or more items each, with either a sufficiently large number of negative covariances and/or at least a few sufficiently large covariances with negative signs among the between-cluster item covariances. Consider Table 4a. Notice that persons A to C pass Items 1 to 4 but fail Item 5, whereas F to J who pass Item 5 tend to fail Items 1 to 4. As a consequence, Item 5's covariances (in the last column of Table 4b) with all the other items are large and negative in sign. The sum of the positive covariances (all of them within the cluster of Items 1 to 4) is 2(0,40) = 0,80, and the sum of the negative covariances is 2(-0,50) = -1,00, so that the total sum of covariances ($2\Sigma_j\Sigma_k s_{jk}$) is -0,20. The total test variance is 1,05 and coefficient alpha turns out to have a negative sign: [5/4][-0,20/1,05] = -0,238.

TABLE 4
AN EXAMPLE OF A NEGATIVE COEFFICIENT ALPHA

4.a: Item matrix

	1	2	3	4	5
A	1	1	1	1	0
B	1	1	1	1	0
C	1	1	1	1	0
D	1	0	0	1	0
E	1	0	0	0	0
F	0	1	0	0	1
G	0	1	0	0	1
H	0	0	1	0	1
I	0	0	1	0	1
J	0	0	0	1	1

4.b: Variance-covariance matrix

	1	2	3	4	5
1	**0,25**	0,05	0,05	0,15	-0,25
2		**0,25**	0,05	0,05	-0,05
3			**0,25**	0,05	-0,05
4				**0,25**	-0,15
5					**0,25**

Most of the authors contemplating the occurrence of negative values for coefficient alpha seem to be oblivious of Cronbach and Hartman's (1954) observation cited above and appear to be at a loss to explain this phenomenon. Henson (2001, p.186) calls such a negative coefficient "a mathematical artifact". Without elaborating, Thompson (2003, p.13) observes that "in practice such a result may mean either that the scores are quite unacceptable, or alternatively that the wrong measurement model has been used to estimate reliability". However, a good starting-point to look for an explanation of negative alphas is the possibility of one or more negatively formulated items that are in need of recoding. (In other words, there may be an error in the scoring key, i.e., "no" responses rather than "yes" responses should earn a mark.) If the scoring of Item 5 in Table 4 is reversed, all its negative covariances become positive and coefficient alpha changes from -0,238 to +0,738. By the same token, if Item 1 in Henson's Table 3 (p. 185) is reversed, the value of coefficient alpha in that example becomes 0,997 instead of -0,373; if the scoring of Items 2, 4 and 6 in Thompson's Table 1.1 (p. 13) is reversed, coefficient alpha changes from -7,00 (notice the absolute value greater than 1,00) to +1,00. Henson (2001, p. 186) says that when the sum of the item covariances is negative "the items seem to be measuring different constructs!" (exclamation mark in the original). Of course, negative item covariances with large absolute values suggest items that measure the opposite poles of the same construct rather than items that are measuring different constructs. (Items with very low covariances irrespective of sign may be measuring different constructs.)

Admittedly, in actual examples the decision which items should have their scoring keys reversed, or whether the scoring key is really the problem at all, may not be as clear-cut as in the present examples. At the same time, items that covary negatively with a large proportion of the other items will probably be detectable by their negative item-total correlations, by their negative loadings in a component or factor analysis, or by a negative sign for their discriminating parameters in latent trait analysis.

ALTERNATIVE COEFFICIENTS

In some situations, the different components of a composite test are of unequal lengths, such as reading passages that contain different numbers of questions. To keep the passages intact, a total score is determined for each passage but due to the passages' varying lengths, individual's true scores on them may no longer be equal (as in parallel components), or differ by the same additive constant (essential tau-equivalence). For situations like these, the assumption of congeneric equivalence, which is less restrictive than essential tau-equivalence, may be appropriate. Whereas the true scores of individuals on essentially tau-equivalent components differ by the same additive constant, in the case of congeneric equivalent components they differ by the same multiplicative constant as well: $T_{ij} = (b)T_{ik} + c$. So if $b = 1{,}5$, the true scores of one individual on two congeneric equivalent components may be 4 and 6, whereas for another it may be 6 and 9 (assuming $c = 0$).

Often the test items are grouped into different subtests in terms of content or difficulty. Under these circumstances, stratified alpha may be more appropriate than coefficient alpha. Its formula is as follows:

$$\text{Stratified alpha} = 1 - [\Sigma\sigma^2_j(1 - \alpha_j)]/\sigma^2_X,$$

where σ^2_j and α_j are the variance and coefficient alpha, respectively, for the jth subtest. When the correlations between items in the same subtest are higher than the correlations across items in different subtests, stratified alpha provides a better estimate than coefficient alpha (Osburn, 2000). For the data in Table 3, stratified alpha turns out to be 0,844 (granted that these data probably would be more consistent with a multidimensional model).

Osburn (2000) performed a simulation study in which various internal-consistency estimates apart from coefficient alpha were computed. In some of the data sets, the components were congeneric equivalent or multidimensional rather than essentially tau-equivalent (which differs from essential tau-equivalence in that the additive constant equals zero). These included an index proposed by Gilmer and Feldt and a maximised adaptation of Guttman's lamda4 which is based on split halves of a test. The six data sets investigated included unidimensional data sets that meet the requirements of parallelism, tau-equivalence and congeneric equivalence, respectively, and two-factor data sets that show increasing degrees of heterogeneity. When the components (unidimensional data) were parallel or tau-equivalent, all the internal-consistency estimates were equal to the true reliability coefficient. For the case in which the components were congeneric equivalent, only Feldt-Gilmer and maximised lambda4 were equal to the true reliability of 0,786, whereas coefficient alpha provided an underestimate, namely, one of 0,778. However, for the two-factor data with different clusters representing factors that were correlated at 0,80, 0,40 and 0,20, only maximised lambda4 was equal to the true reliabilities of 0,781, 0,760 and 0,703, respectively. All other coefficients underestimated the true reliabilities and this underestimation increased as the data became more heterogeneous. For example, both coefficient alpha and Feldt-Gilmer returned values of 0,752, 0,696 and 0,547, respectively, which corresponded to 96,29%, 91,58% and 77,81%, respectively, of the true reliabilities for the three between-factor correlation sizes.

Next, Osburn (2000) grouped four components into two subtests of two components each. Stratified alpha equalled the true reliability irrespective of whether the components were unidimensional (parallel, tau-equivalent or congeneric) and irrespective of the degree of heterogeneity in the two-factor data. By contrast, for the two-factor data coefficient alpha underestimated the true reliability and this underestimation worsened considerably as heterogeneity increased. In the most heterogeneous case (correlation of only 0,20 between the two factors), coefficient alpha was only 0,204 as opposed to a true reliability of 0,613, thus demonstrating this coefficient's inappropriateness in situations in which the assumption of essential tau-equivalence is clearly violated.

DISCUSSION

Several authors (e.g., Green et al., 1977; Hattie, 1985; McDonald, 1981) have lamented the interchangeable use of the terms item *homogeneity*, *internal consistency* and *unidimensionality*. Hattie regards the use of *homogeneity* as a synonym for *unidimensionality* (as suggested by Green et al.) to be redundant, although McDonald would contend that one collection of items may be more homogeneous than another whereas a similar statement cannot be formulated in terms of unidimensionality. All would agree, however, that unidimensionality implies internal consistency as reflected by coefficient alpha but that a high alpha value does not necessarily imply unidimensionality. As a result, an initial dimensionality analysis should be performed if a test or its respective substests are to be interpreted in terms of single dimensions. By the same token, negatively framed items should be recoded following such analyses so that all items that are subjected to internal-consistency analyses are reflective of the same direction.

When items fail to meet the assumption of essential tau-equivalence, more appropriate internal-consistency coefficients than coefficient alpha should be considered, such as the Feldt-Gilmer index and maximised lamda4 in the case of congeneric equivalent data and maximised lamda4 for multidimensional data. There may be no easy way of telling whether the components at hand are essentially tau-equivalent, congeneric equivalent or multidimensional. However, the total scores over subtests of, say, a personality questionnaire designed to measure distinct traits, obviously cannot be regarded as being essentially tau-equivalent.

No degree of diligent test construction can compensate for coefficient alpha's inability to register transient error and hence correct for its tendency to overestimate reliability when such error is present. Just as the potential consequences of ignoring such error were dawning on the psychometric community (cf. Becker, 2000), along came Green (2003) and reinvented, so to speak, coefficient alpha. His test-retest alpha requires that the same test be administered on different occasions and uses the covariances of each item administered on one occasion with every other item administered on the other occasion. In this manner he obtains an index similar to the original coefficient alpha but one that is also susceptible to transient error. As the covariance between any particular item administered on different occasions is ignored in the calculation of the test-retest alpha, it also reduces the possibility of the memory effects that plague the test-retest estimation method. However, the chances that Green's test-retest alpha will reach the same levels of popularity as did the original coefficient alpha are rather slim. The very reason for the original coefficient's popularity has been that it doesn't require a retest, and Green's coefficient has to forgo this luxury, as any other index that wishes to reflect transient error by definition has to do.

In reflecting on his 1951 article in the months prior to his death in 2001, Cronbach (2004) professed his embarrassment about references to *Cronbach's alpha*. Cronbach's (1951) influential

article explicated the meaning of a previously existing quantity. Amongst others, it provided the version that is amenable to multi-point data and which Cronbach labelled coefficient alpha. However, Cronbach (2004) concluded that he no longer considered "the alpha formula as the most appropriate way to examine most data" (p. 14) and pointed out that it provides for "only a small perspective of the range of measurement uses for which reliability information is needed" (p. 29). Certainly, a proper assessment of Cronbach's legacy will have to focus on his role in the development of generalisability theory, an approach in terms of which coefficient alpha is recognised to be reflective of only one source of measurement error among many.

REFERENCES

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5* (3), 370-379.

Bollen, K.A. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, *78* (1), 98-104.

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

Cronbach, L.J. (2004). *My current thoughts on coefficient alpha and successor procedures* (CSE Report 643). Los Angeles, Ca.: University of California.

Cronbach, L.J. & Hartmann, W. (1954). A note on negative reliabilities. *Educational and Psychological Measurement*, *14*, 324-346.

Feldt, L.S. & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Gerbing, D.W. & Anderson, J.C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, *25*, 186-192.

Green, S.B. (2003). A coefficient alpha for test-retest data. *Psychological Methods*, *8,* 88-101.

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

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9* (2), 139-164.

Henson, R.K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Hogan, T.P., Benjamin, A. & Brezinski, K.L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523-531.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error. *Applied Psychological Measurement*, *21*, 337-348.

Lucke, J.F. (2005). The α and ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, *29* (1), 65-81.

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

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

McDonald, R.P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

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

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5* (3), 343-355.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8* (4), 350-353.

Streiner, D.L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80* (1), 99-103.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In Thompson, B. (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-23). Thousand Oaks, Ca.: Sage.