Journal of Applied Psychology
1993, Vol. 78, No. 1, 98-104

Copyright 1993 by the American Psychological Association, Inc.
0021-9010/93/$3.00

What Is Coefficient Alpha? An Examination of Theory and Applications

Jose M. Cortina

Psychological research involving scale construction has been hindered considerably by a widespread lack of understanding of coefficient alpha and reliability theory in general. A discussion of the assumptions and meaning of coefficient alpha is presented. This discussion is followed by a demonstration of the effects of test length and dimensionality on alpha by calculating the statistic for hypothetical tests with varying numbers of items, numbers of orthogonal dimensions, and average item intercorrelations. Recommendations for the proper use of coefficient alpha are offered.

Coefficient alpha (Cronbach, 1951) is certainly one of the most important and pervasive statistics in research involving test construction and use. A review of the Social Sciences Citations Index for the literature from 1966 to 1990 revealed that Cronbach's (1951) article had been cited approximately 60 times per year and in a total of 278 different journals. In addition to the areas of psychology in which one may expect to see alpha used, such as educational, industrial, social, clinical, child, community, and abnormal psychology, this list of journals included representatives from experimental psychology, sociology, statistics, medicine, counseling, nursing, economics, political science, criminology, gerontology, broadcasting, anthropology, and accounting. In spite of its widespread use, however, there is some confusion as to the true meaning and proper interpretation of the statistic.

In this article I address this confusion in two ways. First, a theoretical discussion of alpha is presented. This includes some of the many statements that have been made about alpha and an attempt to integrate these statements. Second, I take a more practical approach in which the interpretation of alpha is examined by observing the changes in alpha as the number of items and interitem correlations are manipulated.

Forms of Reliability

Nunnally (1967) defined reliability as "the extent to which [measurements] are repeatable and that any random influence which tends to make measurements different from occasion to occasion is a source of measurement error" (p. 206). Nunnally went on to explain that there are many factors that can prevent measurements from being repeated perfectly. Although alpha is sometimes referred to as "the" estimate of reliability, it is not the only estimate of reliability. The particular estimate of reliability that one may use depends on the particular error-producing factors that one seeks to identify (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). This is the essence of generalizability theory (Cronbach et al., 1972), which is probably the most widely accepted formulation of reliability. Although a detailed explanation of the theory is not presented here (see Cardinet, Tourneur, & Allal, 1976, or Katerberg, Smith, & Hoy, 1977, for brief explanations of generalizability theory), the basic idea is that aspects of tests or scales (e.g., items, subjects, and raters) are sampled from a predefined domain. Test or scale variance can be broken down into variance attributable to each of these aspects and the interactions among them. The estimate of reliability that one uses must depend on the sources of variance that one considers relevant. If error factors associated with the passing of time are of interest, then test–retest or multiple administrations of parallel tests may be used. If error factors associated with the use of different items are of interest, then internal consistency estimates, such as coefficient alpha (which takes into account variance attributable to subjects and variance attributable to the interaction between subjects and items), or single administrations of parallel tests may be used. Coefficient alpha is obviously not the only estimate of reliability and is inappropriate and insufficient in many cases.

Integrating the Various Descriptions of Alpha

To provide proper interpretations of alpha or any other statistic, one must first understand its meaning. The literature offers many different descriptions of coefficient alpha. Some of these descriptions are contradictory to other such descriptions, some are not, and although I cannot clear up all of these difficulties, some attempt to integrate these views can be made.

Given the variety of perspectives in the literature, perhaps the way to start is to extract those statements about alpha that are commonly accepted in the literature. There seem to be five such statements. They are as follows: (a) Alpha is the mean of all split-half reliabilities (Cronbach, 1951). (b) Alpha is the lower bound of reliability of a test (Kristoff, 1974; Novick & Lewis, 1967; Ten Berge & Zegers, 1978). (c) Alpha is a measure of first-factor saturation (Crano & Brewer, 1973; Hattie, 1985). (d) Alpha is equal to reliability in conditions of essential tau-equivalence (Kristoff, 1974; Novick & Lewis, 1967; Ten Berge & Zegers, 1978). (e) Alpha is a more general version of the Kuder-

Jose M. Cortina, Department of Psychology, Michigan State University.

This article could not have been completed without the guidance and support of many people, most notably, Neal Schmitt, John Hollenbeck, and Stephen Gilliland. I also offer my gratitude to two anonymous reviewers for their thorough and excellent reviews.

Correspondence concerning this article should be addressed to Jose Cortina, Department of Psychology, Michigan State University, East Lansing, MI 48824.

Richardson coefficient of equivalence (Cronbach, 1951; Fiske, 1966; Hakstian & Whalen, 1976). Before I attempt to integrate these statements about alpha, it may be useful to explain them a bit further.

The first statement says that alpha is the mean of all split-half reliabilities. The truth or falsehood of this statement depends on the way one chooses to define both split-half reliability and coefficient alpha. To understand this statement one must understand the distinction between alpha as defined by Cronbach and standardized item alpha. The formula for Cronbach's alpha is

$$\frac{N^2 \times M(\text{COV})}{\text{SUM (VAR/COV)}} \tag{1}$$

where N^2 is the square of the number of items in the scale, M(COV) is the mean interitem covariance, and SUM (VAR/COV) equals the sum of all of the elements in the variance/covariance matrix. The formula for standardized item alpha is essentially the same. The difference is that, for standardized item alpha, the average interitem correlation replaces the average covariance and the sum of the correlation matrix (with ones on the diagonal) replaces the sum of the variance/covariance matrix. This means that Cronbach's alpha takes into account differences in the item standard deviations and is smaller than standardized item alpha to the extent that these differences exist. Standardized alpha is appropriate if item standard scores are summed to form scale scores. Standardized alpha is not appropriate, however, when one chooses to use the simple raw score total as the score for an instrument because, in such a total, differences in item variance affect the total score. In this case, item standard deviations are relevant to the estimation of internal consistency.

Spearman (1910) and Brown (1910) have defined the split-half reliability as the correlation between two halves of a test (r_{12}), corrected to full test length by the Spearman-Brown prophesy formula. Their formula for split-half reliability (r_{sh}) is:

$$r_{sh} = \frac{2r_{12}}{1 + r_{12}}. \tag{2}$$

If this is the definition of split-half reliability that is to be used, then Cronbach's alpha is equal to the mean of all split-half reliabilities only if the item standard deviations are equal. Cronbach's alpha is smaller than this average split-half reliability to the extent that there are differences in item standard deviations.

Flanagan (1937) and Rulon (1939) have given a different formulation of split-half reliability that does take into account standard deviations. Their formula is:

$$r_{sh} = \frac{(4r_{12} \times s_1 \times s_2)}{s_{T^2}}, \tag{3}$$

where s_1 and s_2 are the standard deviations of the halves and s_{T^2} is the variance of the total test. If this is the definition of split-half reliability that is used, then Cronbach's alpha is equal to the average split-half reliability. Formal proof of the equivalence of these two versions of reliability can be found in various textbooks (e.g., Allen & Yen, 1979; Lord & Novick, 1968) as well as Cronbach's (1951) orginal article. Standardized item alpha, which is essentially the average interitem correlation stepped up with the Spearman-Brown formula, does not equal the average split-half reliability from either formulation (except in the extremely unlikely case where the correlations between split halves equal the average interitem correlation), although it is closer to the average Spearman-Brown split half than to the Flanagan-Rulon split half.

Conceptually, the first statement ("Alpha is the mean of all split-half reliabilities.") implies that coefficient alpha (however it is calculated) is a stable estimate of split-half reliability because there is a substantial randomness component to any estimate of split-half reliability. This randomness stems from the fact that any estimate of split-half reliability that one gets depends, to some extent, on the particular manner in which one chooses to split the test. Cronbach's alpha, because it is the mean of all possible splits (as measured by Flanagan and Rulon), is not subject to this randomness and is therefore more stable.

The second and fourth statements, which deal with alpha as a lower bound of reliability that is equal to reliability under tau-equivalence, are related to each other and are described in more detail later.

The third statement says that alpha is a measure of first-factor saturation. This statement suggests that alpha is a measure of the extent to which there is a general factor present in a set of items and, therefore, the extent to which the items are interrelated. This statement, however, contradicts what was said by Cronbach in the original article, and it has been shown to be false with respect to Cronbach's alpha by subsequent research. This research and its implications for the third statement are discussed later. What is important is that the third statement is true, at least to some extent, with respect to standardized item alpha (Kaiser, 1968). Kaiser (1968) showed that, if all item intercorrelations were equal to the average item intercorrelation (i.e., the set of items has exactly one principal component), then standardized alpha is directly related to the eigenvalue of the first unrotated principal component. Because this relationship depends on unidimensionality, standardized alpha is inappropriate to the extent that more than one factor is responsible for the correlations among a set of items.

What Does Alpha Measure?

The fifth statement says that alpha is a general version of the Kuder-Richardson coefficient of equivalence. It is a general version because the Kuder-Richardson coefficient applies only to dichotomous items, whereas alpha applies to any set of items regardless of the response scale. This fact is explained thoroughly in Cronbach's (1951) article and need not be repeated here, but the description of alpha as a coefficient of equivalence does lead to another important issue.

In Cronbach's (1947) classic description of different types of reliability, test variance was depicted as a sum of general, group, and specific variance and changes in each (if the test is repeated) as well as a residual term. For a coefficient of equivalence, *error* is defined as the variance due to specific factors and a residual term (Cronbach, 1947). For some operationalizations of reliability, some or all of the variance associated with group factors may contribute to error. This is the point that Hattie

(1985) and Cureton (1958) were trying to make. Alpha is said to be a measure of first-factor saturation (i.e., the extent to which a certain factor is present in all items), and although a set of items with a high alpha usually has a strong common factor, this is not necessarily the case. Green, Lissitz, and Mulaik (1977), in a Monte Carlo study, generated data corresponding to a 10-item test that occupied a five-dimensional common factor space. Each item loaded equally (.45) on two of five orthogonal factors, no two items loaded on the same two common factors, and each item had a communality of .90. The alpha for this set of items was calculated to be .81. By commonly accepted interpretation, this would be taken as an indication of unidimensionality when it should not be. Instead, it is an indication that the set of items conforms to Cronbach's (1951) definition of equivalence, which is that there is very little variance specific to individual items. All 10 items loaded high on more than one of the factors; there was very little item-specific variance. The utility of information about the amount of item-specific variance in a test is described in the Discussion.

Another way to approach this issue of dimensionality is to examine the confusion in the literature of the terms *internal consistency* and *homogeneity*. Some authors have failed to make a distinction between the two terms (e.g., Nunnally, 1970) when it seems that a distinction needs to be made. *Internal consistency* refers to the degree of interrelatedness among the items (Crano & Brewer, 1973; Green et al., 1977), whereas *homogeneity* refers to unidimensionality (Green et al., 1977; Gulliksen, 1950; Lord & Novick, 1968). As Green et al. (1977) pointed out, internal consistency is a necessary but not sufficient condition for homogeneity. Alpha is, among other things, a function of internal consistency, that is, of interrelatedness of items. A set of items, however, can be relatively interrelated and multidimensional. The Monte Carlo study by Green et al. (1977) pointed this out. Alpha was high in spite of the fact that one third of the item intercorrelations were zero. So, one conclusion that can be drawn with respect to what alpha measures is this:

> It is a function of the extent to which items in a test have high communalities and thus low uniquenesses. It is also a function of interrelatedness, although one must remember that this does not imply unidimensionality or homogeneity.

Precision of Alpha

It was mentioned above that a set of items can be somewhat interrelated and multidimensional. This is not so much an issue for the level of alpha, but rather for the precision of alpha. As Nunnally (1978) explained, "To the extent that correlations among items in a domain vary, there is some error connected with the average correlation found in any particular sampling of items" (p. 206).

Precision is measured in terms of the standard error of item intercorrelations, which, in turn, is a function of the variance of the item intercorrelations. This is clearly evident in the formula for the standard error of alpha:

$$\frac{SD_r}{[(1/2 \times k \times [k-1]) - 1]^{1/2}}, \quad (4)$$

where SD_r is the standard deviation of item intercorrelations and k is the number of items. The level of alpha, as I show earlier, is a function of the size of the average correlation among items and can be large in spite of a wide range of item intercorrelations. The precision of alpha, because it is a function of the spread of item correlations, reflects this range of correlations regardless of the source or sources of the range (e.g., measurement error or multidimensionality). For example, examine the item intercorrelation matrices for two 4-item tests in Table 1.

In spite of the fact that these two intercorrelation matrices are radically different from one another, standardized alpha for each of these two sets of items is .63. The difference is reflected in the standard error or precision of the estimate of reliability. The precision estimate for the first matrix is obviously zero. The estimate of reliability is the same no matter which of these items one uses to calculate it. The precision estimate for the second matrix, however, is .13. Given the heterogeneity of the second set of items, the estimate of reliability varies greatly depending on which items are chosen to represent this domain and to estimate alpha in a particular instance. Also, the intercorrelations in the second matrix suggest that the set of items is composed of two dimensions. A large standard error, although it does not provide enough information by itself to prove multidimensionality, is a symptom of multidimensionality. The resulting implications for interpreting alpha are discussed later.

A final implication of the earlier quotation from Nunnally (1978) is that the assumptions that one makes about how items are to be sampled from a domain affects the estimate of reliability. For example, standardized alpha for both matrices in Table 1 is .63. However, parallel forms reliability (Lord & Novick, 1968, p. 98), which assumes overall parallelism of composites (instead of random sampling from the domain assumed by alpha), yields a higher estimate of reliability in the heterogenous case (.84) than it does in the homogenous case (.63). The reason for this is that parallel forms reliability, in essence, considers the interrelatedness of sets of items within each factor, whereas alpha considers the interrelatedness of the total set of items. Because the correlations among items within factors is higher in the heterogenous case than in the homogenous case, parallel forms reliability is higher in the heterogenous case. It

Table 1
Intercorrelation Matrices for Two Sets of Items With Different Standard Errors of Reliability

	No. of items		
No. of items	1	2	3
	Precision estimate = 0		
1	—		
2	.30	—	
3	.30	.30	—
4	.30	.30	.30
	Precision estimate = .13		
1	—		
2	.70	—	
3	.10	.10	—
4	.10	.10	.70

must be made clear, however, that parallel forms reliability is only appropriate when composite parallelism (not item-for-item parallelism) is a reasonable assumption. That is, parallel forms are constructed so that the multiple factors represented in the domain of items are represented equally in both parallel forms.

Alpha as a Reliability Estimate

If one combines the second and fourth statements ("Alpha is the lower bound of reliability of a test." and "Alpha is equal to reliability in conditions of essential tau-equivalence."), then one sees that Cronbach's alpha is a lower bound of reliability and that it approaches reliability as the measurements become essentially tau-equivalent. Measurements are essentially tau-equivalent if they are linearly related and differ only by a constant. Cronbach's alpha is a lower bound of reliability because perfect essential tau-equivalence is seldom if ever achieved (standardized alpha is not a lower bound, but is a direct approximation of reliability given items with equal observed variance). When tests comprise equal portions of general and group factor variance (in Cronbach's, 1947, model), then their items are essentially tau-equivalent, and Cronbach's alpha equals reliability. So, a second statement that we can make about alpha is this:

> As the items in tests approach essential tau-equivalence (i.e., linearly related and differing only by a constant), as they do when the tests are composed of equal portions of general and group factor variance, Cronbach's alpha approaches reliability. When test items are exactly essentially tau-equivalent, Cronbach's alpha equals reliability.

Current Usages of Alpha

Another lesson to be learned from Green et al.'s (1977) Monte Carlo study is that alpha (either Cronbach's or standardized) is a function of the number of items in a scale. Although most who use alpha pay lip-service to this fact, it seems to be forgotten when interpreting alpha. Most recent studies that have used alpha imply that a given level, perhaps greater than .70, is adequate or inadequate without comparing it with the number of items in the scale. Any perusal of the recent literature in applied psychology supports this statement. This acceptance of $\alpha > .70$ as adequate is implied by the fact that $\alpha > .70$ usually goes uninterpreted. It is merely presented, and further scale modifications are seldom made. This is clearly an improper usage of the statistic. As an example, I compare the meaning of standardized $\alpha = .80$ for scales made up of 3 and 10 items.

For a 3-item scale with $\alpha = .80$, the average interitem correlation is .57. For a 10-item scale with $\alpha = .80$, the average interitem correlation is only .28. This is strikingly different from .57 and underscores the fact that, even without taking dimensionality into account, alpha must be interpreted with some caution. This is not to say that the absolute level of alpha is meaningless. The proportion of error variance for a test or scale with $\alpha = .80$ is exactly the same for any test regardless of the number of items. What one must keep in mind when evaluating test or scale characteristics is that, for example, 40 items (any 40 items if one assumes they are not correlated zero or negatively with each other) has a relatively large alpha simply because of the number of items, and number of items is, to say the least, an inadequate measure of test or scale quality. This is not a criticism of alpha per se. As I said, alpha is a sound measure of proportion of error variance regardless of test length. This simply suggests that when many items are pooled, internal consistency estimates are relatively invariant (i.e., large) and therefore somewhat useless.

One reason for the misuse of alpha in applied psychology is that there seems to be no real metric for judging the adequacy of the statistic. Experience with the literature gives one some general idea of what an acceptable alpha is, but there is usually little else to go on. Note, however, that those who make decisions about the adequacy of a scale on the basis of nothing more than the level of alpha are missing the point of empirically estimating reliability. The level of reliability that is adequate depends on the decision that is made with the scale. The finer the distinction that needs to be made, the better the reliability must be. For example, the reliability of the Scholastic Aptitude Test is quite adequate for distinguishing between a 750 scorer and a 450 scorer. Its reliability is not adequate for distinctions between scores of 749 and 750. Thus, any judgment of adequacy, even in research, needs to consider context (J. Hollenbeck, personal communication, June 19, 1991).

Number of Items, Dimensionality, and Alpha

It has been said that alpha is appropriately computed only when there is a single common factor (Cotton, Campbell, & Malone, 1957). If there is only one common factor, then alpha is a measure of the strength of that factor. The problem is that, just as the psychological literature reflects no clear understanding of the extent to which alpha is affected by the number of items, so does it reflect no clear understanding of the extent to which alpha is affected by dimensionality.

Green et al. (1977) gave some notion of the extent to which alpha changes as a function of the number of items. From a practical standpoint, however, there are two problems in the interpretation of their results. First, they manipulated several aspects of sets of items other than number of items, such as number of factors that determine items and communalities. This makes it difficult to assess the effects of any one variable. Second, Green et al. manipulated numbers of items between 19 and 75. This, in turn, causes two problems: (a) The relationship between number of items and alpha is curvilinear (Komorita & Graham, 1965) and begins to level off before the number of items reaches 19, and (b) many if not most of the scales that are used in applied research today have fewer than 19 items.

The fact that the numbers of items in the data created by Green et al. (1977) are so large also hinders those who generally use somewhat shorter scales from seeing the extent to which the number of items in a scale hides the dimensionality of a scale as measured by alpha. What may be useful is a display of the effects of changes in dimensionality and number of items on coefficient alpha. This is precisely my next goal. Specifically, alpha was calculated for scales with different numbers of items, different numbers of orthogonal dimensions, and different average item intercorrelations. Specifically, alphas were calculated for scales with 1, 2, and 3 dimensions, 6, 12, and 18 items,

and average item intercorrelations of .30, .50, and .70. Before I present these data, two clarifications are important.

First, dimensions within scales were made orthogonal in the interest of simplicity. Although few, if any, dimensions in reality are completely orthogonal, especially within the same scale, any level of dimension intercorrelation for this study would have been chosen completely arbitrarily. Also, nonzero dimension intercorrelations only serve to strengthen my claims (i.e., that alpha alone is not a measure of unidimensionality). When there are, in fact, completely independent dimensions, alpha is at its lowest.

Second, *average item intercorrelation* refers to the average item intercorrelation within each dimension of a scale. So, for example, an average item intercorrelation of .50 in the one dimension condition means that the average item intercorrelation for the entire scale is .50, whereas the same average item intercorrelation in the two-dimension condition means that the average within each of the two dimensions is .50. The average for the entire scale is smaller because items across dimensions were made orthogonal and, therefore, correlated zero.

Results

Alphas for the conditions described earlier are presented in Table 2.

The first three rows in Table 2 display alphas for unidimensional scales and suggests three conclusions. First, number of items has a profound effect on alpha, especially at low levels of average item intercorrelation. Second, in a unidimensional scale an average item intercorrelation of .50 yields alphas that are acceptable by convention (i.e., greater than .75) regardless of number of items. Third, if a scale has enough items (i.e., more than 20), then it can have an alpha of greater than .70 even when the correlation among items is very small. In short, dimensionality notwithstanding, alpha is very much a function of the number of items in a scale, and although alpha is also a function of item intercorrelation, it must be interpreted with number of items in mind. Alpha gives us information about the extent to which each item in a set of items correlates with at least one other item in the set (i.e., the communalities of the items). Such groups of items within a set are usually more interpretable than items that stand alone in a correlation matrix. (This notion is explained in more detail later.) Therefore, alpha offers important information, but one must keep in mind the fact that alpha does not offer information about other types of error, such as error associated with stability over time (e.g., changes in ability, practice, or memory).

The relationship between test length and reliability has long been known and is the subject of graphs in many psychometric texts (e.g., Lord & Novick, 1968). The first three rows in Table 2 simply show the somewhat surprisingly large extent to which this is true. Perhaps less known, or at least less appreciated, are the results displayed in the parts of Table 2 labeled "Two dimensions" and "Three dimensions."

The second set of rows displays alphas and precision estimates for scales with two dimensions and, besides adding support to the conclusions drawn from the first set of rows, suggests three conclusions with respect to dimensionality.

First, if a scale has more than 14 items, then it will have an alpha of .70 or better even if it consists of two orthogonal dimensions with modest (i.e., .30) item intercorrelations. If the dimensions are correlated with each other, as they usually are, then alpha is even greater. Second, if the two dimensions themselves have high average item intercorrelations (i.e., $r > .70$), then alpha can be quite good (i.e., greater than .85). Third, the precision of alpha (the standard error of the correlations in the item intercorrelation matrix) offers far more information about

Table 2
Alphas and Precision Estimates for Scales With Different Numbers of Dimensions, Different Numbers of Items, and Varying Average Intercorrelations

	Average item intercorrelation					
	$r = .30$		$r = .50$		$r = .70$	
No. of items	α	Precision	α	Precision	α	Precision
			One dimension			
6	.72		.86		.93	
12	.84		.92		.96	
18	.88		.95		.98	
			Two dimensions			
6	.45	.04	.60	.07	.70	.09
12	.65	.02	.78	.03	.85	.04
18	.75	.01	.85	.02	.90	.03
			Three dimensions			
6	.28	.03	.40	.05	.49	.08
12	.52	.02	.65	.03	.74	.04
18	.64	.01	.76	.02	.84	.02

Note. Because the scales with one dimension are absolutely unidimensional, precision = 0 for all of them.

dimensionality than the size of alpha. For example, as indicated in the second set of rows, coefficient alpha is .85 for a test with 12 items in the instance of a correlation of .70. This sizable alpha, however, does not say anything about unidimensionality. Instead, the estimate of precision (.04) suggests that there is a departure from unidimensionality. Also, it is important to remember that although the sizable alpha gives important information about the communalities of the items, it does not give information about stability across time.

The third set of rows displays alphas and precision estimates for scales with three dimensions and serves to strengthen the points made with respect to the second set of rows.

Specifically, the third set shows that, given a sufficient number of items, a scale can have a reasonable alpha even if it contains three orthogonal dimensions. If the three dimensions themselves have average item intercorrelations of .70 or better and are somewhat correlated with each other, then the overall scale has an alpha of .80 or better. Again, any attempts to draw conclusions about dimensionality are better served by an estimate of precision. For example, in the 12-item case with the correlation of .70, coefficient alpha is .74. The precision of alpha, however, is .04, which reflects the extent to which the correlations in the item intercorrelations matrix differ from each other.

One final note about this table is that although the purpose of presenting this table is to show that alpha can be high in spite of low item intercorrelations and multidimensionality, alpha does increase as a function of item intercorrelation, and alpha does decrease as a function of multidimensionality. The lesson to be learned from this table is that alpha can be high in spite of low item intercorrelations and multidimensionality.

Discussion

The purpose of this article is to explain how coefficient alpha has been misunderstood in applied psychological research and to show how alpha is affected by the number of items, item intercorrelations, and dimensionality. In an attempt to promote appropriate use of alpha, I provide an explanation of the original assumptions of the statistic as a coefficient of equivalence and the limitations on interpretation that these assumptions impose.

I demonstrate the extent to which alpha is affected by the number of items, average item intercorrelation, and dimensionality by calculating alphas for scales with three different numbers of dimensions, levels of average item intercorrelation, and numbers of items. These calculations show that alpha can be rather high and acceptable by the standards of many (greater than .70) in spite of low average item intercorrelation or multidimensionality, provided there is a sufficient number of items. Although most researchers who use coefficient alpha are aware of these issues, the results presented in Table 2 point out the surprising range of alphas that are possible with increases in the number of items as well as the surprising size of alphas that are possible even with pronounced multidimensionality.

It may be argued that multidimensionality is irrelevant because if a test has a good alpha, then it is free of error associated with the use of different items, just as a test with good test–retest reliability is free of error associated with the passing of time. This is certainly true, but it does not mean that the total score on a multidimensional test has a straightforward or unambiguous interpretation. An adequate coefficient alpha (number of items notwithstanding) suggests only that, on the average, split halves of the test are highly correlated. It says nothing about the extent to which the two halves are measuring the construct or constructs that they are intended to measure. Even if the total score of a test could perhaps be used for some practical purpose like selection, it could not be interpreted. In other words, the test would be known to measure something consistently, but what that is would still be unknown. Some form of construct validation is necessary to establish the meaning of the measure.

When Is Alpha Useful?

Although the results of this article may be interpreted as pessimistic toward the usefulness of coefficient alpha, this was not intended. Instead, the purpose was to remind those who construct tests and need to use some measure of internal consistency that alpha is not a panacea. Just like any other statistic, it must be used with caution. Coefficient alpha is useful for estimating reliability in a particular case: when item-specific variance in a unidimensional test is of interest. If a test has a large alpha, then it can be concluded that a large portion of the variance in the test is attributable to general and group factors. This is important information because it implies that there is very little item-specific variance. These concepts come from Cronbach (1947) and are analogous to factor-analytic terms. For example, consider a standardized test like the Graduate Record Examination. The general factor for such a test may be the reading component that is present in all of the items. The group factors may be Verbal, Quantitative, and Analytical. Any additional variance is item-specific. This item-specific variance is called the *uniqueness* of the item, and it is this uniqueness that is assessed with coefficient alpha. Again, this is important information because we can usually interpret general and group factors (as has been done with the Graduate Record Examination) but not item-specific variance. The problem with interpretation arises when large alpha is taken to mean that the test is unidimensional.

One solution to such problems with the statistic is to use one of the many factor-analytic techniques currently available to make sure that there are no large departures from unidimensionality. For example, the first step in establishing unidimensionality is to conduct a principal-components analysis of a set of items. This provides information similar to that provided by the estimate of precision. If this analysis suggests the existence of only one factor, then alpha can be used to conclude that the set of items is unidimensional. The principal-components analysis alone does not provide enough evidence to conclude that a set of items is unidimensional because such an analysis may, for example, yield only one factor even if the items have correlations of .10 with each other. In essence, what this means is that alpha can be used as a confirmatory measure of unidimensionality or as a measure of the strength of a dimension once the existence of a single factor has been determined. As always, the number of items must be kept in mind. For these purposes, Table 2 or similar tables and graphs in Lord and Novick (1968)

can be used as a referent so that one can plug in the number of items in a scale and its average intercorrelation or its alpha to get an idea of the extent to which an alpha of, for example, .70 really does reflect internal consistency instead of irrelevancies like the number of items.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterrey, CA: Brooks/Cole.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13,* 119–136.

Cotton, J. W., Campbell, D. T., & Malone, R. D. (1957). The relationship between factorial composition of test items and measures of test reliability. *Psychometrika, 22,* 347–358.

Crano, W. D., & Brewer, M. B. (1973). *Principles of research in social psychology.* New York: McGraw-Hill.

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.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory for generalizeability of scores and profiles.* New York: Wiley.

Cureton, E. E. (1958). The definition and estimation of test reliability. *Educational and Psychological Measurement, 18,* 715–738.

Fiske, D. W. (1966). Some hypotheses concerning test adequacy. *Educational and Psychological Measurement, 26,* 69–88.

Flanagan, J. C. (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology, 28,* 17–21.

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.

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika, 15,* 259–269.

Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219–232.

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

Kaiser, H. F. (1968). A measure of the average intercorrelation. *Educational and Psychological Measurement, 28,* 245–247.

Katerberg, R., Smith, F. J., & Hoy, S. (1977). Language, time, and person effects on attitude translation. *Journal of Applied Psychology, 62,* 385–391.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25,* 987–995.

Kristoff, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39,* 491–499.

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

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

Nunnally, J. C. (1967). *Psychometric theory* (1st ed.). New York: McGraw-Hill.

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

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

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

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3,* 271–295.

Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43,* 575–579.

Received December 2, 1991
Revision received May 18, 1992
Accepted May 18, 1992 ■