# Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha

Robin K. Henson

*Although often ignored, reliability is critical when interpreting study effects and test results. Accordingly, this article focuses on the most commonly used estimate of reliability, internal consistency coefficients, with emphasis on coefficient alpha. An interpretive framework is provided for applied researchers and others seeking a conceptual understanding of these estimates.*

Reliability is often a misunderstood measurement concept. Nevertheless, the reliability of a test's scores bears on clinical decisions regarding diagnosis and treatment as well as on the findings and interpretations of counseling and development researchers. There are a variety of forms of reliability coefficients, but among the most commonly used are internal consistency estimates because they are readily calculated from a single administration of a test. Hogan, Benjamin, and Brezinski (2000) found that about 75% of reported reliability estimates in the *Directory of Unpublished Experimental Mental Measures* (published by the American Psychological Association [APA]) were internal consistency estimates. A quick perusal of the research literature would verify the popularity of these coefficients. Cronbach's (1951) prophecy that his internal consistency coefficient is "a tool that we expect to become increasingly prominent in the research literature" (p. 299) has clearly come to pass. It is important to understand, however, that internal consistency coefficients are not direct measures of reliability, but rather are theoretical estimates derived from classical test theory. More direct (although still ultimately theoretical) measures of reliability include test–retest and alternate forms coefficients, which address score stability across time or consistency between test forms, respectively.

Internal consistency estimates relate to item homogeneity, or the degree to which the items on a test jointly measure the same construct. Whenever a test's items are linearly combined into a single composite score, as is often the case in clinical and research usage, the issue of item homogeneity speaks directly to the ability of the clinician or the researcher to interpret the composite score as a reflection of all the test's items. Counseling and development researchers often use tests to assess individual differences on a variety of constructs, and authors of articles in *Measurement and Evaluation in Counseling and Development* commonly examine the psychometric properties of their assessments (see, e.g., Loo, 2001, on

*Robin K. Henson is an assistant professor of educational research in the Department of Technology and Cognition at the University of North Texas, Denton. An earlier version of this article was presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY, November 2000. Correspondence regarding this article should be sent to Robin K. Henson, Department of Technology and Cognition, University of North Texas, PO Box 311337, 1300 Highland, Denton, TX 76203-1337 (e-mail: rhenson@tac.coe.unt.edu).*

motivation toward work; McCarthy, Moller, & Fouladi, 2001, on parental attachment and affect regulation; and Pajares, Hartley, & Valiante, 2001, on writing self-efficacy). Internal consistency reliabilities are important in both substantive and measurement contexts. As Elmore, Ekstrom, and Diamond (1993) noted, "Without a confident knowledge of basic concepts such as error of measurement, counselors may use test results inappropriately" (p. 123). Tymofievich and Leroux (2000) suggested, therefore, that the client may incur "the negative consequences not only of a poorly developed test but also of a counselor who is unable to recognize it as such" (p. 50).

The importance of reliability is further reflected in a recently published report by the APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999). The Task Force recommended that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (p. 596). The recommendation to report score reliability in all studies stems directly from the Task Force's mandate to "always present effect sizes for primary outcomes" (p. 599) because "interpreting the size of observed effects requires an assessment of the reliability of the scores" (p. 596).

Effect size magnitude is inherently attenuated by the reliability of the scores used to obtain the effect estimate (Reinhardt, 1996). As Reinhardt observed,

> Reliability is critical in detecting effects in substantive research. For example, if a dependent variable is measured such that the scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results will not be statistically significant at any sample size, including an incredibly large one. (p. 3)

Despite their importance, empirical studies document that effect sizes are seldom reported (cf. Henson & Smith, 2000; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000). In a similar manner, reliability estimates are more often than not overlooked or omitted in published articles (cf. Henson, Kogan, & Vacha-Haase, 2001; Vacha-Haase, Ness, Nilsson, & Reetz, 1999; Yin & Fan, 2000). Articles reporting effects in light of reliability are almost nonexistent.

The reliability of the scores in any study, measurement and substantive, is central to understanding the observed relationships between variables. Because all classical analyses (e.g., *t* test, analysis of variance, regression, canonical correlation) are part of the same general linear model and are correlational in nature (Bagozzi, Fornell, & Larcker, 1981; Cohen, 1968; Henson, 2000; Knapp, 1978; Thompson, 1991), most studies should report and interpret results in light of the reliability for the present data, which usually means consulting an internal consistency estimate (Thompson, 1994).

Unfortunately, too few researchers report score reliability for their studies, and even fewer interpret their effects in light of reliability. This deficit in the literature is likely due to myriad factors, chief of which is the common misconception that reliability inures to tests, rather than to scores (cf. Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998). A contrary view is given by Sawilowsky (2000a, 2000b). Because scores may vary in degree of reliability, a given test may yield grossly divergent reliability estimates on different administrations. The reader is referred to Caruso (2000), Henson et al. (2001), Viswesvaran and Ones (2000), Yin and Fan (2000), and Vacha-Haase (1998) for examples of this phenomenon.

Different samples, testing conditions, and any other factor that may affect observed scores can in turn affect reliability estimates. Because reliability inherently attenuates effect sizes, it can also affect statistical power, an often overlooked point (Onwuegbuzie & Daniel, 2000). Because effects and power may be attenuated by the reliability of observed scores, reliability should always be reported and considered in result interpretation (Wilkinson & APA Task Force on Statistical Inference, 1999).

## PURPOSE

Pedhazur and Schmelkin (1991) suggested that many researchers' misconceptions and unawareness surrounding reliability may be due to decreased emphasis on measurement course work in doctoral programs. Aiken et al. (1990) verified this measurement vacuum in doctoral curricula. In a national survey of American Educational Research Association members, Mittag and Thompson (2000) found less than desirable understanding of reliability among respondents.

Accordingly, the present article is intended to provide an interpretive framework for applied researchers and others seeking a conceptual understanding of internal consistency reliability estimates. Internal consistency estimates are discussed here (rather than other coefficients such as test–retest and interrater reliability) because of their frequency of use and ease of calculation in most research situations. This article (a) briefly reviews some basic tenets of classical test theory; (b) discusses the salient factors that affect internal consistency reliability estimates, with emphasis on coefficient alpha; and (c) presents several suggestions for improving understanding (and use) of score reliability.

## SOME BASIC TENETS OF CLASSICAL TEST THEORY

Reliability is concerned with score consistency and is particularly relevant when there are important ramifications of our interpretations. The more measurement error that exists in our scores, the less useful these scores may be for analysis, interpretation, and clinical purposes. This section addresses several key, albeit brief, points related to the classical test theory underlying internal consistency estimates. The reader is referred to Crocker and Algina (1986) for a complete treatment. Important historical discussions include those by Gulliksen (1950), Lord and Novick (1968), Nunnally (1967, 1978), Stanley (1971), and Thorndike (1951). Treatment of the limitations inherent in classical test theory have been given by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and Kieffer (1999).

### Ratio of Score Variances: The General Linear Model in Measurement

The classical conceptualization of score reliability relates the concept of score consistency to "true scores." A person's true score is the theoretical average obtained from an infinite number of "independent testings of the same person with the same test" (Allen & Yen, 1979, p. 57). In other words, for any measurement occasion that is less than perfect, an obtained set of scores will contain variance that is true score variance (measuring the trait of interest) and variance that is due to error (factors inhibiting trait measurement, e.g., chance in selection of answers by sheer guessing). The sum of these two variances yields the total score variance of the observed scores, such that

$$s_{\mathrm{TRUE}}{}^{2} + s_{\mathrm{ERROR}}{}^{2} = s_{\mathrm{TOTAL}}{}^{2} \qquad (1)$$

For example, if 20% of a test's total score variance is attributable to nonsystematic error, coefficient alpha would be .80 (this statistic is discussed in detail later), indicating that 80% of the total score variance is reliable. In classical test theory, systematic errors (e.g., consistent fatigue effect across the sample vs. random fatigue effects) are not considered measurement error and actually help increase the reliability estimate.

Equation 1 makes the attenuation of reliability on effect sizes explicit, because only true score variance may be correlated between any two variables (or linear composite sets of variables beyond the bivariate case). It is impossible to correlate random error across variables, thereby attenuating an $r^2$ type effect size to be less than 1. This phenomenon can also be observed by squaring both sides of the commonly known "correction for attenuation" formula, such that

$$r_{xy}{}^2 = r_{\mathrm{TxT}y}{}^2(r_{xx})(r_{yy}) \qquad (2)$$

where $r_{\mathrm{TxT}y}{}^2$ is the squared correlation between the "true" scores of the variables $x$ and $y$ (i.e., the "corrected" effect) and $r_{xx}$ and $r_{yy}$ are the internal consistency reliabilities of the variables $x$ and $y$, respectively. From this formula, $r_{xy}{}^2$ will be less than or equal to the product of the reliabilities (Nunnally, 1967, 1978; Nunnally & Bernstein, 1994). For example, if one variable has an alpha = .70 and another variable has an alpha = .60, the maximum possible effect would be (.70)(.60) = .42 = $r_{xy}{}^2$. It can also be observed from Equation 2 that the $r_{xy}{}^2$ effect will reach its theoretical maximum only when (a) the reliabilities of the variables are perfect and (b) $r_{\mathrm{TxT}y}{}^2 = 1$, that is, when the correlation between the "true" scores of the variables is perfect.

Another generalization of Equation 1 informs us that reliability can be conceptualized as a ratio of true score variance to total (observed) score variance (80% in the above example). Dawson (1999) noted that coefficient alpha is an analog of the more familiar $r^2$ type effect, and, accordingly, represents a ratio of variances. Dawson generalized the $r^2$ statistic and noted that any univariate $r^2$ type variance-accounted-for statistic (e.g., $r^2$, $R^2$, $\eta^2$) can be computed as a ratio of variance explained ($\sigma_{\mathrm{EXPLAINED}}{}^2$) divided by the total variability in the dependent variable ($\sigma_{\mathrm{TOTAL}}{}^2$). Conceptually, this statistic asks, "What portion (or percentage) of the total information [in the dependent variable] can an extraneous variable explain or predict?" (Dawson, 1999, p. 105).

For coefficient alpha (a common internal consistency reliability estimate; Cronbach, 1951), this same ratio of variances is apparent in the following formula:

$$\mathrm{a} = k/(k - 1)\ [1 - (\mathrm{S}s_k{}^2/s_{\mathrm{TOTAL}}{}^2)] \qquad (3)$$

where $k$ = the number of items on the test, $\Sigma\sigma_k{}^2$ = the sum of all the $k$ item variances, and $\sigma_{\mathrm{TOTAL}}{}^2$ = the variance of the total test scores. In the alpha formula, the ratio of variances is captured in the ($\Sigma\sigma_k{}^2/\sigma_{\mathrm{TOTAL}}{}^2$) term.

Because of this ratio of variances, Dawson (1999) noted that the general linear model that guides much substantive statistical analysis also infuses the measurement context: "The presence of the general linear model (GLM) across both substantive and measurement analyses can also be seen in the computation of coefficient alpha (Cronbach, 1951) as the ratio of two variances" (p. 109). However, as Thompson (1999) noted, "Psychometrically alpha involves more than only variances and their ratios to each other" (p. 12). Most explicitly, the alpha formula invokes $\Sigma\sigma_k{}^2$ as the numerator, which is related to, but different from, the variance explained as noted by Dawson (this issue is explained momentarily along with an illustration of coefficient alpha).

## Internal Consistency Reliability Estimates

Typically, many authors conceptualize three sources of measurement error within the classical framework: content sampling of items, stability across time, and interrater error (see, e.g., Anastasi & Urbina, 1997; Hopkins, 1998; Popham, 2000). Content sampling refers to the theoretical idea that the test is made up of a random sampling of all possible items that could be on the test. If so, the items should be highly interrelated because they assess the same construct of interest (e.g., self-esteem, achievement). This item interrelationship is typically called "internal consistency," which suggests that the items on a measure should correlate highly with each other if they truly represent appropriate content sampling. If items are highly correlated, it is theoretically assumed that the construct of interest has been measured to some degree of consistency (i.e., the scores are reliable).

Because internal consistency estimates are intended to apply to test items assumed to represent a single underlying construct, the use of these estimates with speeded tests is inappropriate because of the confounding of construct measurement with testing speed. Furthermore, for tests that consist of scales measuring different constructs, internal consistency should be assessed separately for each scale.

So, exactly how large does an internal consistency estimate need to be before one can consider scores to be reliable? Of course, the exact magnitude of the estimate will vary depending on the purposes of the research and uses of the scores. Nevertheless, Nunnally (1967) noted that in "the early stages of research on predictor tests or hypothesized measures of a construct, . . . reliabilities of .60 or .50 will suffice" (p. 226). For basic research purposes, Nunnally suggested .80. For applied settings in which cutoff scores may be important (e.g., special education placement, college admission tests), .90 is the minimally tolerable estimate, with .95 as "the desired standard" (Nunnally, 1967, p. 226). In Nunnally's second edition of his classic work (1978), the exploratory standard for instrument development was increased to .70, resulting in many researchers citing Nunnally (1978) if they attained this loftier standard and citing the first edition if they did not! These standards were held constant in the most recent edition (Nunnally & Bernstein, 1994). Accordingly, as Loo (2001) noted in a recent *Measurement and Evaluation in Counseling and Development* article, internal consistency estimates are often considered to have a "generally accepted .80 cutoff value" (p. 223) for general research purposes. Standardized test scores used for important clinical and/or educational decisions should have reliabilities of .90 or better (Hopkins, 1998; Nunnally & Bernstein, 1994; Oosterhof, 2001).

As a measure of internal consistency and a generalization of the older split-half method, Kuder and Richardson (1937) presented their classic formula, KR-20 (named such because the formula was the 20th listed in their article), as follows:

$$\text{KR-20} = k/(k-1)[1-(\Sigma p_k q_k/\sigma_{\text{TOTAL}}^{2})] \qquad (4)$$

where $k$ = the number of items on the test, $p_k$ = the proportion of people correctly answering item $k$, $q_k$ = the proportion of people incorrectly answering item $k$ (i.e., $1 - p_k$), and $\sigma_{\text{TOTAL}}^2$ = the variance of the total test scores. Because $\Sigma p_k q_k$ deals with mutually exclusive proportions for two possible outcomes, it should be clear that KR-20 works only when test items are dichotomously scored (e.g., 0 vs. 1). This formula may apply to either achievement or attitude measures, as long as scoring is dichotomous (e.g., correct vs. incorrect, agree vs. disagree).

It is important to note that the variance of a dichotomously scored item ($\sigma_k^2$) will always equal $p_k q_k$. If all persons responded the same way to an item, then $\sigma_k^2 = p_k q_k = 0$, because no variance would be present in the scores. Furthermore, if one half of the responses were scored 0 and the other half scored 1, then the scores would have maximum variability. When items are dichotomously scored, the maximum variability possible is $\sigma_k^2 = p_k q_k = .25$. This is because each squared deviation score will be .25, a result of subtracting the mean of .5 from 0 or 1 and squaring this difference. The sum of these squared deviation scores (i.e., sum of squares) divided by $n$ (variance) will result in .25, regardless of sample size (cf. Reinhardt, 1996).

Fourteen years after the advent of KR-20, Cronbach (1951) introduced coefficient alpha, a more general form of the KR-20 formula. Coefficient alpha is defined in Equation 3. Comparison of the KR-20 and alpha formulas reveals that only the numerator of the variance ratio differs. Because $\Sigma\sigma_k^2 = \Sigma p_k q_k$ as noted above, it should be apparent that alpha can be used with dichotomously scored items. However, because the sum of the item variances is used as the numerator, alpha can also be used with measures using multiple response categories, such as Likert scale data.

### The Importance of Total Score Variance

In both KR-20 and alpha, it is clear that certain data features will lead to higher reliability estimates. Specifically, by holding the number of items constant ($k$), reliability will increase as the sum of item variances decreases and the total score variance increases.

Accordingly, classical reliability estimates hinge on the variance of the total scores. Total variance not only is relevant for test–retest and alternate forms reliability but it is also a critical component of internal consistency estimates. As this variance increases, the reliability estimate will also tend to increase, due to greater *theoretical* confidence that we have appropriately ordered (measured) the participants on the trait of interest (Henson et al., 2001). One implication of the role of total score variance is that different samples will likely yield different reliabilities because the total variance will likely change. For example, Thompson (1994) observed, "The same measure, when administered to more heterogeneous or more homogeneous sets of subjects, will yield scores with differing reliability" (p. 839).

It is important that the various classical test theory sources of measurement error (e.g., internal consistency, test–retest, interrater) are separate and cumulative (Anastasi & Urbina, 1997). Too many researchers believe that if they obtain $\alpha = .90$ for their scores, then the same 10% of measurement error would be found in a test–retest or interrater coefficient. Instead, assuming 10% error for internal consistency, stability, and interrater, then the overall measurement error would be 30%, not 10%, because these estimates explain *different sources* of error. As an aside, generalizability theory allows for the simultaneous examination of these sources of error as well as the *interactions* between them by using analysis of variance methodology. Readers are referred to Shavelson and Webb (1991) for an accessible treatment of generalizability theory.

## A CONCEPTUAL PRIMER ON COEFFICIENT ALPHA

As noted, alpha invokes a general linear model ratio of explained variance to total variance as a fundamental component in its calculation. However, as a measure of internal consistency, it also must account for the intercorrelation among the items, with the assumption that as items are more highly correlated the magnitude of alpha will increase.

Three heuristic examples are used here to illustrate the salient data features that affect coefficient alpha. These examples are heavily dependent on Thompson (1999) and Reinhardt (1996) and were adapted for use here.

### Example 1: Perfectly Uncorrelated Items

Although test items typically are correlated to some degree, the present example illustrates the impact on alpha when items are perfectly uncorrelated ($r$ and covariance = 0 for all pairwise item combinations). Table 1 presents a heuristic data set for four test items with interitem correlations of 0.

On basis of the above Equation 3 for alpha, reliability can be estimated if we can identify the number of items, the sum of the item variances, and the variance of the total scores. The first two of these items is given in Table 1, with $k = 4$ and the sum of the item variances as $.73 = (.22 + .18 + .18 + .15)$. Crocker and Algina (1986, p. 95) presented a formula for the calculation of the total score variance using only the Table 1 data:

$$\sigma_{\mathrm{TOTAL}}{}^2 = \Sigma\sigma_k{}^2 + [\Sigma \mathrm{COV}_{ij} \text{ (for } i < j) \times 2] \qquad (5)$$

Close examination of this formula reveals that total test score variance can be conceptualized as an additive function of two components: (a) the sum of the item variances ($\Sigma\sigma_k{}^2$) and (b) the doubled sum of the unique covariances [$\Sigma \mathrm{COV}_{ij}$ (for $i < j$) × 2]. This formula

## TABLE 1

**Example 1: Item Correlations (Covariances) Are Zero and Calculation of Total Test Score Variance**

| | Correlation | | | | Variance/Covariance | | | |
|---|---|---|---|---|---|---|---|---|
| **Variance** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 1 | — | | | | $\underline{.22}$ | | | |
| 2 | .00 | — | | | .00 | $\underline{.18}$ | | |
| 3 | .00 | .00 | — | | .00 | .00 | $\underline{.18}$ | |
| 4 | .00 | .00 | .00 | — | .00 | .00 | .00 | $\underline{.15}$ |

| **Pairing** | COV/Variance | | | $r/\sigma$ | | |
|---|---|---|---|---|---|---|
| $i < j$ | $COV_{ij}$ | $\sigma_i^2$ | $\sigma_j^2$ | $r_{ij}$ | $\sigma_i$ | $\sigma_j$ |
| 1 2 | .00 | .22 | .18 | .00 | .47 | .42 |
| 1 3 | .00 | .22 | .18 | .00 | .47 | .42 |
| 1 4 | .00 | .22 | .15 | .00 | .47 | .39 |
| 2 3 | .00 | .18 | .18 | .00 | .42 | .42 |
| 2 4 | .00 | .18 | .15 | .00 | .42 | .39 |
| 3 4 | .00 | .18 | .15 | .00 | .42 | .39 |

*Note.* $COV_{ij}$ represents the recalculated covariance using $COV_{ij} = r_{ij}(\sigma_i \times \sigma_j)$. These estimates match the original covariances ($COV_{ij}$) and illustrate the $r$ to COV transformation. Underlined values represent the item's variance. $\Sigma COV_{ij} = .00$. $\Sigma COV_{ij} \times 2 = .00$.

highlights the important point that the total test score variance is at least partially dependent on the intercorrelations among the items on a test, a finding in harmony with the idea that alpha is a measure of internal consistency. Table 1 presents calculations for determining the covariance portion of the total test score variance. Table 1 also illustrates the COV to $r$ transformation, thereby demonstrating the relationship between these important statistics. Using the data from Table 1, the total test score variance is found with

$$\begin{aligned}\sigma_{\text{TOTAL}}^2 &= \Sigma\sigma_k^2 + [\Sigma COV_{ij} \text{ (for } i < j) \times 2] \\ &= (.22 + .18 + .18 + .15) + .00 \\ &= .73\end{aligned}$$

These calculations indicate that in this example, the total score variance is *only* a function of the sum of the individual item variances, because the covariances were 0. This finding verifies that total score variance ($\sigma_{\text{TOTAL}}^2$) will equal the sum of item variances ($\Sigma\sigma_k^2$ or $\Sigma p_k q_k$) "only when the covariances among items are 0" (Sax, 1974, p. 182).

Now using the total score variance as our last remaining piece of information, alpha can be found with

$$\begin{aligned}\alpha &= k/(k - 1)\ [1 - (\Sigma\sigma_k^2/\sigma_{\text{TOTAL}}^2)] \\ &= 4/(4 - 1)\ [1 - (.22 + .18 + .18 + .15)/.73] \\ &= 4/3\ [1 - (.73/.73)] \\ &= 1.33\ [1 - 1] \\ &= 1.33\ [0] \\ &= 0\end{aligned}$$

Because the items shared no variance, such that the covariances and correlations were 0, it stands to reason that there was no internal consistency among the items. Accordingly, alpha's calculations led to this logical conclusion ($\alpha = 0$). Furthermore, on the basis of this understanding, the alpha formula reveals that we should expect alpha to increase as the covariances contribute more to the total score variance.

### Example 2: Perfectly Correlated Items

When items are perfectly correlated, and thereby possess perfect internal consistency, we should no doubt expect alpha to reach its maximum of 1 (representing 100% of true score variance due to content sampling). Table 2 presents data on four perfectly correlated test items. Table 2 also presents the calculations necessary to obtain the total score variance using Equation 5. Using these results, the total score variance is

$$\begin{aligned}\sigma_{\text{TOTAL}}^{2} &= \Sigma\sigma_{k}^{2} + [\Sigma\text{COV}_{ij}\ (\text{for } i < j) \times 2]\\ &= (.22 + .18 + .18 + .15) + (1.08 \times 2)\\ &= .73 + 2.16\\ &= 2.89\end{aligned}$$

Using the total score variance, alpha is

$$\begin{aligned}\alpha &= k/(k-1)\ [1 - (\Sigma\sigma_{k}^{2}/\sigma_{\text{TOTAL}}^{2})]\\ &= 4/(4-1)\ [1 - (.22 + .18 + .18 + .15)/2.89]\\ &= 4/3\ [1 - .73/2.89]\\ &= 1.33\ [1 - .2525952]\\ &= 1.33\ [.7474048]\\ &= .9940\end{aligned}$$

As expected, $\alpha = 1$ (within rounding error due to calculation of the covariances in Table 2), indicating perfect internal consistency of scores.

### Example 3: Perfectly Correlated Items With Mixed Signs

It is possible for items to be highly correlated but not all in the same direction. Table 3 presents the heuristic data matrices for perfectly correlated items with mixed signs and presents calculations that lead to the total score variance. The total score variance is

$$\begin{aligned}\sigma_{\text{TOTAL}}^{2} &= \Sigma\sigma_{k}^{2} + [\Sigma\text{COV}_{ij}\ (\text{for } i < j) \times 2]\\ &= (.22 + .18 + .18 + .15) + (-.08 \times 2)\\ &= .73 + (-.16)\\ &= .57\end{aligned}$$

Coefficient alpha is solved as

$$\begin{aligned}\alpha &= k/(k-1)\quad [1 - (\Sigma\sigma_{k}^{2}/\sigma_{\text{TOTAL}}^{2})]\\ &= 4/(4-1)\ [1 - (.22 + .18 + .18 + .15)/.57]\\ &= 4/3[1 - .73/.57]\\ &= 1.33[1 - 1.2807018]\\ &= 1.33[-.2807018]\\ &= -.3733\end{aligned}$$

Here we have found what Thompson (1999, p. 15) called a "paradox" in the calculation of alpha. That is, how can alpha be *negative,* given that it is a squared metric statistic ($r^2$

## TABLE 2

### Example 2: Item Correlations Are One and Calculation of Total Test Score Variance

| | Correlation | | | | Variance/Covariance | | | |
|---|---|---|---|---|---|---|---|---|
| Variance | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | — | | | | .22 | | | |
| 2 | 1.00 | — | | | .20 | .18 | | |
| 3 | 1.00 | 1.00 | — | | .20 | .18 | .18 | |
| 4 | 1.00 | 1.00 | 1.00 | — | .18 | .16 | .16 | .15 |

| Pairing | COV/Variance | | | r/σ | | |
|---|---|---|---|---|---|---|
| $i < j$ | $\text{COV}_{ij}$ | $\sigma_i^2$ | $\sigma_j^2$ | $r_{ij}$ | $\sigma_i$ | $\sigma_j$ |
| 1 2 | .20 | .22 | .18 | 1.00 | .47 | .42 |
| 1 3 | .20 | .22 | .18 | 1.00 | .47 | .42 |
| 1 4 | .18 | .22 | .15 | 1.00 | .47 | .39 |
| 2 3 | .18 | .18 | .18 | 1.00 | .42 | .42 |
| 2 4 | .16 | .18 | .15 | 1.00 | .42 | .39 |
| 3 4 | .16 | .18 | .15 | 1.00 | .42 | .39 |

*Note.* See Table 1 *Note.* $\Sigma\text{COV}_{ij} = 1.08$. $\Sigma\text{COV}_{ij} \times 2 = 2.16$.

type ratio of variances)! Solving for alpha with the equivalent formula presented by Sax (1974, p. 181) helps provide a deeper understanding of alpha's ratio of variances:

$$\begin{aligned}\alpha &= k/(k-1)\ [(\boldsymbol{\sigma}_{\mathbf{TOTAL}}{}^2 - \boldsymbol{\Sigma\sigma}_k{}^2)/\text{s}_{\text{TOTAL}}{}^2)] \\ &= 4/(4-1)\ [(\mathbf{.57} - \mathbf{.73})/.57] \\ &= 4/3\ [\mathbf{-.16}/.57] \\ &= 1.33\ [-.2807018] \\ &= -.3733\end{aligned} \tag{6}$$

## TABLE 3

### Example 3: Item Correlations With Mixed Signs and Calculation of Total Test Score Variance

| | Correlation | | | | Variance/Covariance | | | |
|---|---|---|---|---|---|---|---|---|
| Variance | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | — | | | | .22 | | | |
| 2 | −1.00 | — | | | −.20 | .18 | | |
| 3 | −1.00 | 1.00 | — | | −.20 | .18 | .18 | |
| 4 | −1.00 | 1.00 | 1.00 | — | −.18 | .16 | .16 | .15 |

| Pairing | COV/Variance | | | r/σ | | |
|---|---|---|---|---|---|---|
| $i < j$ | $\text{COV}_{ij}$ | $\sigma_i^2$ | $\sigma_j^2$ | $r_{ij}$ | $\sigma_i$ | $\sigma_j$ |
| 1 2 | −.20 | .22 | .18 | −1.00 | .47 | .42 |
| 1 3 | −.20 | .22 | .18 | −1.00 | .47 | .42 |
| 1 4 | −.18 | .22 | .15 | −1.00 | .47 | .39 |
| 2 3 | .18 | .18 | .18 | 1.00 | .42 | .42 |
| 2 4 | .16 | .18 | .15 | 1.00 | .42 | .39 |
| 3 4 | .16 | .18 | .15 | 1.00 | .42 | .39 |

*Note.* See Table 1 *Note.* $\Sigma\text{COV}_{ij} = -.08$. $\Sigma\text{COV}_{ij} \times 2 = -.16$.

Here we find that the numerator essentially represents the covariances between the test items, which follows from Equation 5 used to calculate the total score variance. In the numerator of Equation 6, we have essentially removed the sum of the item variances ($\Sigma\sigma_k^2$) from the total score variance ($\sigma_{\text{TOTAL}}^2$), which leaves the doubled sum of item covariances [$\Sigma\text{COV}_{ij}$ (for $i < j$) × 2]. The covariance term is found in the boldface calculations for alpha above (**–.16**) and in the calculations in Table 3. Thus, the alpha ratio essentially includes the sum of the item *covariances* over the total score *variance*. Accordingly, we would expect alpha to increase when the item correlations are large *and* in the same direction (see Equation 5).

The negative result ($\alpha = -.37$) that we find in the present example, then, is a mathematical artifact that occurs when the sum of the item variances exceeds the total score variance. Conceptually, this would mean that the individual variability of the *k* items tends to be greater than the shared variability (covariance/correlation) between the *k* items. If this is true, then internal consistency suffers because the items seem to be measuring different constructs! In keeping with a classical test theory perspective, the psychometric properties of alpha (and KR-20) capture this conceptual expectation. Of course, from a practical standpoint, a negative alpha is theoretically impossible and essentially represents zero reliability. For this reason, the reliability estimate should be reported as zero rather than as the negative coefficient.

## TOWARD A BETTER UNDERSTANDING (AND USE) OF SCORE RELIABILITY

### Reliability Affects Power

Reliability inherently attenuates the maximum possible magnitude of relationships between variables (see previous discussion in this article of attenuation of effect size). Accordingly, all else being constant, poor score reliability will reduce the power of statistical significance tests (cf. Onwuegbuzie & Daniel, 2000). When effects are reduced, they become harder to find. Researchers would be compelled to increase sample size or their level of statistical significance to compensate for this loss of power.

When researchers find nonstatistically significant results due to poor measurement, the bottom-line ramifications may include greater difficulty publishing in a literature biased toward statistically significant results, the tendency to ignore potentially meaningful effects, and a perpetuated misunderstanding of why the results were not statistically significant (i.e., ignoring a potential measurement problem). Therefore, interpretation of effects warrants examination of the observed reliability (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596). For a more complete discussion of statistical significance tests, the reader is referred to the seminal work of Cohen (1990, 1994) as well as Henson and Smith (2000) and Thompson (1994, 1996).

### Reporting Practices and Interpretation

Researchers should report reliability for their scores, and not depend on estimates from prior studies or test manuals. As correctly noted by Gronlund and Linn (1990), "Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself. Thus, it is more appropriate to speak of the reliability of 'test scores' or the 'measurement' than of the 'test' or the 'instrument' " (p. 78). Furthermore, researchers would do well to use precise language when referencing the reliability of their scores.

Unfortunately, *empirical* studies confirm that very few researchers actually report reliability estimates for their data (cf. Caruso, 2000; Vacha-Haase, 1998; Yin & Fan, 2000). Because so few researchers report reliability, and even fewer interpret effects in light of reliability, the practical impact of this effect attenuation is largely unknown. As researchers report reliability for the data in hand, and consider these estimates when interpreting their results, more will be learned about reliability's impact on power, effect sizes, and statistical significance tests.

## Reliability Generalization Studies

Because reliability varies on different administrations of a test, Vacha-Haase (1998) used a meta-analytic method called "reliability generalization" (RG) that allows examination of the variability of score reliability across studies. In addition, coded study characteristics (e.g., sample composition and variability) can be used as potential predictors of reliability variation, thereby providing some evidence of which sampling conditions most affect score reliability. Vacha-Haase's method is based on the older validity generalization approach (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977) and represents an important development in the examination of score integrity. A primary benefit of RG studies is the cumulative information they may yield in describing study characteristics that affect reliability estimates for scores from a given test and, perhaps, study characteristics that consistently affect score reliability across different tests.

Reporting of reliability coefficients would not only inform the study in which the reliability was reported but also facilitate meta-analytic RG studies. Readers are referred to Henson and Thompson (2001), Vacha-Haase (1998), and Thompson and Vacha-Haase (2000) for more complete discussions of RG.

In sum, score reliability is important in both measurement and substantive studies as well as clinical application of measures. Therefore, researchers ought to report reliability for their data and interpret results in light of the obtained estimates. This practice would move the field toward a better understanding and use of reliability estimates.

## REFERENCES

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., with Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). The training in statistics, methodology, and measurement in psychology. *American Psychologist, 45,* 721–734.

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Bagozzi, R. P., Fornell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16,* 437–454.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60,* 236–254.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Chicago: Holt, Rinehart & Winston.

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 of generalizability for scores and profiles.* New York: Wiley.

Dawson, T. E. (1999). Relating variance portioning in measurement analyses to the exact same process in substantive analyses. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 101–110). Stamford, CT: JAI Press.

Elmore, P. B., Ekstrom, R., & Diamond, E. E. (1993). Counselors' test use practices: Indicators on the adequacy of measurement training. *Measurement and Evaluation in Counseling and Development, 26,* 116–124.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Henson, R. K. (2000). Demystifying parametric analyses: Illustrating canonical correlation as the multivariate general linear model. *Multiple Linear Regression Viewpoints, 26*(1), 11–19.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61,* 404–420.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA task force report and current trends. *Journal of Research and Development in Education, 33,* 285–296.

Henson, R. K., & Thompson, B. (2001, April). *Characterizing measurement error in test scores across studies: A tutorial on conducting "reliability generalization" analyses.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Kieffer, K. M. (1999). Why generalizability theory is essential and classical test theory is often inadequate. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 149–170). Stamford, CT: JAI Press.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85,* 410–416.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Loo, R. (2001). Motivational orientations toward work: An evaluation of the Work Preference Inventory (Student Form). *Measurement and Evaluation in Counseling and Development, 33,* 222–233.

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

McCarthy, C. J., Moller, N. P., & Fouladi, R. T. (2001). Continued attachment to parents: Its relationship to affect regulation and perceived stress among college students. *Measurement and Evaluation in Counseling and Development, 33,* 198–213.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14–20.

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

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Onwuegbuzie, A. J., & Daniel, L. G. (2000, November). *Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Oosterhof, A. (2001). *Classroom applications of educational measurement* (3rd ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Pajares, F., Hartley, J., & Valiante, G. (2001). Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development, 33,* 214–221.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd ed.). Boston: Allyn & Bacon.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 3–20). Greenwich, CT: JAI Press.

Sawilowsky, S. S. (2000a). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some *EPM* editorial policies. *Educational and Psychological Measurement, 60,* 157–173.

Sawilowsky, S. S. (2000b). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement, 60,* 196–200.

Sax, G. (1974). *Principles of educational measurement and evaluation.* Belmont, CA: Wadsworth.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24,* 80–95.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25,* 26–30.

Thompson, B. (1999, February). *Understanding coefficient alpha, really.* Paper presented at the annual meeting of the Educational Research Exchange, College Station, TX.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174–195.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Tymofievich, M., & Leroux, J. A. (2000). Counselors' competencies in using assessments. *Measurement and Evaluation in Counseling and Development, 33,* 50–59.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6–20.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67,* 335–341.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10,* 413–425.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60,* 224–235.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201–223.