DOCUMENT RESUME

| | |
|---|---|
| ED 448 204 | TM 032 234 |
| AUTHOR | Onwuegbuzie, Anthony J.; Daniel, Larry G. |
| TITLE | Reliability Generalization: The Importance of Considering Sample Specificity, Confident Intervals, and Subgroup Differences. |
| PUB DATE | 2000-11-00 |
| NOTE | 44p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (28th, Bowling Green, KY, November 17-19, 2000). |
| PUB TYPE | Reports - Descriptive (141) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | Correlation; *Generalization; *Reliability; *Research Methodology; Statistical Significance |

ABSTRACT

The purposes of this paper are to identify common errors made by researchers when dealing with reliability coefficients and to outline best practices for reporting and interpreting reliability coefficients. Common errors that researchers make are: (1) stating that the instruments are reliable; (2) incorrectly interpreting correlation coefficients; (3) not reporting reliability coefficients for their own sample; (4) conducting tests of statistical significance on reliability coefficients; and (5) failing to report the reliability of difference scores when examining gain scores. It is recommended that researchers report reliability coefficients for their own data and that they interpret confidence intervals around reliability coefficients, considering that reliability coefficients should be reported not only for the full sample at hand but also for each subgroup. A heuristic example is used for the two-sample case (i.e., t-test) to illustrate how comparing subgroups with different reliability coefficients can affect statistical power. (Contains 3 tables and 36 references.) (Author/SLD)

Running head: RELIABILITY GENERALIZATION

# Reliability Generalization: The Importance of Considering Sample Specificity, Confidence Intervals, and Subgroup Differences

Anthony J. Onwuegbuzie

Valdosta State University

Larry G. Daniel

University of North Florida

Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, Kentucky, November 15-17, 2000.

## Abstract

The purposes of the present paper were to identify common errors made by researchers when dealing with reliability coefficients and to outline best practices for reporting and interpreting reliability coefficients. Common errors that researchers make include (a) stating that instruments are reliable, (b) incorrectly interpreting correlation coefficients, (c) not reporting reliability coefficients for their own sample, (d) conducting tests of statistical significance on reliability coefficients, and (e) failing to report reliability of difference scores when examining gain scores. It is recommended that researchers report reliability coefficients for their own data and that they interpret confidence intervals around reliability coefficients, considering that reliability coefficients represent only point estimates. Further, it is contended that reliability coefficients should be reported not only for the full sample at hand, but also for each subgroup. A heuristic example is utilized for the two-sample case (i.e., *t*-test) to illustrate how comparing subgroups with different reliability coefficients can affect statistical power.

# Reliability Generalization: The Importance of Considering Sample Specificity, Confidence Intervals, and Subgroup Differences

Measurement is the most important component of the research process. For example, even if an extremely large sample is selected utilizing random sampling techniques, and even if sophisticated data-analytical procedures are employed, the underlying research question(s), however important, cannot be adequately addressed if the data collected are not trustworthy. This is true for both quantitative and qualitative research. With respect to the former (i.e., quantitative or "empirical" research), there are two important characteristics that scores from any measuring instrument should possess: *validity* and *reliability*.

## VALIDITY

Validity is to the extent to which scores generated by an instrument measure the trait or variable they are intended to measure for a given population. In other words, validity refers to the appropriateness of the interpretations made from instrument scores with respect to a particular use. The concept of validity is very straightforward when measuring physical constructs or attributes. For example, if one is attempting to measure the time taken by athletes in an Olympic meeting to cross the 100-meter line, then it follows that the times recorded will be valid if the measuring device used (e.g., a stopwatch) is correctly calibrated and used correctly. However, in the field of education, the concept of validity is not as clear. For instance, if a researcher attempts to measure students' level of statistics anxiety by administering a self-report measure, it is considerably more difficult to establish whether scores generated by this instrument truly represent students' level of anxiety. Within the cognitive domain, in which important educational decisions often are made, establishing evidence for validity of test data is especially important.

Establishing Evidence of Validity

As previously noted, validation refers to the process of systematically collecting evidence to justify the array of inferences that are intended to be drawn from scores generated by an instrument (American Educational Research Association, American Psychological Association, & National Council on Measurement and Evaluation [AERA, APA, & NCME], 1985). In validation studies, researchers attempt to obtain one or more of three types of evidences: *content-related validity*, *criterion-related validity*, and *construct-related validity*. Although validity is most appropriately thought of as a unitary concept, the aforementioned types of validity evidence are sometimes erroneously referred to as "validity types" or "categories of validity" (AERA, APA, & NCME, 1985):

> the various means of accumulating validity evidence have been grouped into categories called *content-related, criterion-related* and *construct-related evidence* of *validity*. These categories are convenient, as are other more refined categorizations. . ., but the use of the category labels does not imply that there are distinct types of validity or that a specific validation strategy is best for each specific inference or test use. (p. 9)

*Content-related validity evidence* addresses the extent to which the items on an instrument represent the content to be measured. Establishment of content-related validity evidence includes attention to *face validity* (the extent to which the items appear relevant, important, and interesting to the subject), *item validity* (the extent to which the specific items represent measurement in the intended content area), and *sampling validity* (the extent to which the full set of items samples the total content area). Face validity is often regarded as extremely weak validity evidence as it is based strictly on face value of the items in a test after the test has already been constructed (Nunnally & Bernstein, 1994). Even though item and sampling validity are typically considerations one uses during the development of a test, these approaches to validity evidence are also limited as they rely heavily upon expert judgement, not empirical analysis. Content-related validation studies typically involve several experts who examine an instrument's content systematically and evaluate the extent to which the items represent the content domain adequately. Alternately, a test developer may present content-related validity evidence based on showing similarities between item content and a codified representation of the content domain (e.g., a textbook, a curriculum manual, a body of literature articulating the nature of an educational construct).

*Criterion-related validity evidence* pertains to the extent to which scores on a measuring instrument are related to an independent external variable (i.e., criterion) believed to measure directly the underlying attribute or behavior. Once the external criterion has been operationalized, empirical data are obtained in order to assess the relationship between scores on the measuring instrument and scores on the criterion. The resulting correlation coefficient that indexes this relationship represents one form of a *validity coefficient*. This coefficient indicates how accurately the scores on the measure can predict the criterion. The general term "criterion-related validity" may be used to refer to both *concurrent validity* and *predictive validity*. Concurrent validity measures the degree to which scores on an instrument are related to scores on another, already-established instrument administered (approximately) simultaneously, or to a measurement of some other criterion that is available at the same point in time as the scores on the instrument of interest. Predictive validity refers to the relationship between instrument scores and criterion scores that are measured at a future time.

*Construct-related validity evidence* is based on the accumulation of a number of independent validation studies. The objective in seeking evidence for construct-related validity is to operationalize the underlying construct by means of a theoretical framework, as well as to determine how well it is being measured. Methods used in construct-related validation include (a) defining the domain to be measured and comparing scores of known groups (i.e., *contrasted groups approach*); (b) comparing scores before and after some particular intervention; (c) correlating scores yielded from the instrument of interest with scores from other instruments that measure the same construct (i.e., *convergent validity*); (d) correlating scores generated from the instrument of interest with scores from instruments that measure concepts theoretically and empirically related to but not the same as the construct of interest (i.e., *discriminant validity*); (e) correlating scores yielded from the instrument of interest with measures of constructs antithetical to the construct of interest (i.e., *divergent validity*); (f) collating the results of various independent studies that use the instrument (i.e., *successive verification)*; (g) manipulating a relevant independent variable and observing whether the instrument scores change (i.e., *experimental studies*); (h) studying the intercorrelations among a set of items or instrument scores in order to determine the number of factors (constructs) needed to account for the intercorrelations (i.e., *factor analysis)*; and (i) examining the instrument itself and collecting information about the content of the instrument, the processes used in responding to items on the instrument, and relationships among the items (i.e., *intra-measure analysis).* According to Messick (1981), construct-related evidence is the most comprehensive because it subsumes and extends content-related and criterion-related evidence.

## RELIABILITY

A second major characteristic of test scores, reliability, will be the focus the remainder of this paper. Reliability refers to the extent to which scores yielded by an instrument administered to specific individuals, at a specific point in time, and under certain conditions, are reproducible. For scores to be reproducible, they must be consistent. Thus, the most popular definition of reliability is that it pertains to the extent that scores are consistent, regardless of whether the scores are measuring what they have been designed to measure (i.e., whether the scores are valid). As stated by Crocker and Algina (1986, p. 105), “In practical terms reliability is the degree to which individuals’ deviation scores, or z-scores, remain relatively consistent over repeated administration of the same test or alternate test forms.”

In the physical sciences, many properties of objects can be measured with near-perfect reliability. However, this is not the case for the social sciences in general and for the field of education in particular. Constructs of interest in the social sciences are typically abstractions (e.g., personality, achievement, intelligence, motivation, locus of control) that must be measured indirectly; hence, the vast majority of measures in the social sciences generate scores that are, to some degree, unreliable. For example, if two parallel achievement measures were administered to the same group of students, it would be virtually impossible for each student to obtain identical scores on both measures.

Unreliability occurs as a result of errors of measurement, which can be random or systematic. Random errors of measurement are the result of chance occurrences that stem from factors such as fluctuations in the administration of the instrument, variations in the respondent's mental or psychological state (e.g., levels of alertness and anxiety), guessing, and scoring errors. Systematic errors reflect errors that consistently affect individuals' scores because of a particular characteristic of an individual respondent, the group of respondents, or the instrument that is independent of the underlying construct. Whereas random errors reduce both the consistency and utility of scores, by either inflating or depressing any respondent's score in an unpredictable manner, systematic errors adversely influence the practical usefulness of the scores (Crocker & Algina, 1986).

The concept of reliability stems from the classical true score model, the basis for which was laid by Charles Spearman (1907, 1913). The true score model contains the following three components: observed score (O), true score (T), and random error (E). These components are related in the following manner:

$$O = T \pm E$$

As the formula illustrates, the smaller the error term, the closer the observed score approximates the true score. In fact, the true score would be obtained if there were no error in measurement. Thus, the true score component pertains to scores that an individual would obtain if the instrument yielded perfect measurements of the construct.

In the above model, a positive error results in true score overestimation, whereas a negative error culminates in true score underestimation. Because underestimation and overestimation are equally likely to occur, the average (mean) error is expected to be zero if the same instrument is administered an infinite number of times. Consequently, a true score is the individual's mean score on an infinite number of measurements. Obviously, because it is not possible to administer an instrument an infinite number of times, true score is a theoretical concept. When an instrument is administered, only the observed scores are known although it is the true scores that are really of interest. Thus, it is important that observed scores closely approximate their true scores--or, stated differently, that the observed scores and true scores are highly related. This relationship is measured via the *reliability index*, which can be expressed as the ratio of the standard deviation of true scores to the standard deviation of the observed scores. Similarly, the *reliability coefficient*, which represents the correlation between scores on parallel instruments, is the ratio of the true-score variance to the observed-score variance in a set of scores. Thus, the reliability coefficient is the square of the reliability index (Nunnally & Bernstein, 1994).

## Estimating Score Reliability

Because the true score is a theoretical rather than observed measure, researchers, clinicians, and practitioners who collect test data can, at best, derive estimates of reliability based on the set of scores generated by their data. The four most common methods of estimating a reliability coefficient are (a) administering an instrument to the same group of individuals on two or more occasions and correlating the paired scores (i.e., *test-retest reliability,* or *stability reliability*); (b) administering two different measures of a construct at essentially the same time to the same group of individuals and then correlating the paired scores (i.e., *equivalence reliability,* or *alternate forms reliability*), (c) administering two different measures of a construct at two separate occasions to the same group of individuals and then correlating the paired scores (i.e., *coefficient of stability and equivalence*), and (d) estimating the reliability of scores based on alternate configurations of the items across one administration of the instrument (i.e., *coefficient of internal consistency*). Although, theoretically, the coefficient of reliability ranges from 0 (measurement is all error) to 1 (no error in measurement), the four methods of estimating the reliability coefficient presented above make it possible for the reliability estimate to be negative (although this is rare). Thus, technically, reliability coefficients may range from -1 to 1 even though negative values are intuitively meaningless. (A dataset cannot contain less than 0% true score variance!)

Whereas there are many occasions in which it is either inappropriate or impossible to determine the coefficient of stability (e.g., when the instrument is a developmental measure) or the coefficient of equivalence (e.g., when only one version of the instrument is required), as noted by Crocker and Algina (1986), it is always appropriate to estimate the internal consistency coefficient for scores on an instrument because this coefficient represents an index of both "item content homogeneity and item quality" (p. 135). In any case, the coefficient of stability, coefficient of equivalence, and coefficient of stability and equivalence typically represent underestimates of the theoretical reliability coefficient that would be obtained from truly parallel measures and/or varied administration conditions. Consequently, the remainder of this essay will deal specifically with the most popular method of estimating score reliability, namely, the *coefficient of internal consistency*.

### *Issues Pertaining to Coefficients of Internal Consistency*

The "internal consistency" of scores on an instrument can be estimated in several ways. These methods include split-half methods and procedures based on item covariances. Split half methods involve first dividing the instrument into two artificial subscales (e.g., at random, contrasting odd-numbered items and even-numbered items), with each subscale representing half the length of the original measure. Second, these two subscales are then scored separately for each respondent, and the correlation ($r_{1\cdot2}$) is computed between the two sets of scores. Finally, the Spearman-Brown prophecy (i.e., [2 x $r_{1\cdot2}$] / [1 + $r_{1\cdot2}$]) can be applied to the correlation between the two subscales in order to obtain a corrected estimate of the reliability coefficient for scores on the full-length instrument. Unfortunately, split-half techniques do not yield a unique estimate of score reliability because there are many possible ways of dividing an instrument into two subscales of equal length, with each configuration of items potentially yielding a different reliability estimate. Thus, split-half techniques are not as widely utilized for estimating internal consistency are methods based on item covariances.

The most commonly-used methods of estimating the internal consistency of items are Cronbach's (1951) coefficient alpha and Kuder and Richardson's (1937) KR-20 formula. The former (i.e., coefficient alpha) is computed by the following formula:

$$\hat{\alpha} = \frac{k}{k - 1}\left(1 - \frac{\Sigma\hat{\sigma}_i^2}{\Sigma\hat{\sigma}_x^2}\right)$$

where *k* is the number of items on the instrument, $\Sigma\hat{\alpha}_i^2$ is the sum of the individual item variances, and $\hat{\alpha}_x^2$ is the variance of the total instrument scores across all respondents. This formula is appropriate when all items underlying the instrument are dichotomously scored or when items have a specific number of fixed responses (e.g., Likert-format scales). Alternatively, the KR-20 formula can be utilized as a measure of internal consistency, as follows:

$$KR_{20} = \frac{k}{k - 1}\left(1 - \frac{\Sigma p_i q_i}{\Sigma\hat{\sigma}_x^2}\right)$$

where *k* is the number of items on the instrument, $p_i q_i$ is the sum of the individual item variances, and $\hat{\alpha}_x^2$ is the variance of the total instrument scores across all respondents. This formula is equivalent to coefficient alpha when $\hat{\alpha}_i^2$ can be substituted by $p_i q_i$. However, KR-20 can be used only for dichotomously-scored items. As such, KR-20 is a special case of coefficient alpha or, alternately stated, coefficient alpha is a more general form of KR-20.

Both coefficient alpha and KR-20 are bounded by ±1. As noted by Roberts and Onwuegbuzie (2000), both internal consistency estimates will be large when scores that yield a small sum of individual item variances are associated with a large total score variance. Similarly, scores that yield large individual item variances but a small total score variance will produce small alpha coefficients. Although rare, it is possible to obtain a negative alpha reliability, which occurs when the sum of the individual item variances is larger than the total instrument variance (Roberts & Onwuegbuzie, 2000).

## Errors Made By Researchers Pertaining to Reliability Coefficients

*Incorrect interpretation of reliability coefficients.* As noted by Vacha-Haase, Kogan, and Thompson (in press), many writers use phrases such as "the instrument is reliable." However, it is the scores yielded that are reliable, not the instrument. Additionally, some researchers interpret an instrument that generates a reliability coefficient of, say, .70 for a set of scores, as being 70% reliable. A more correct interpretation of a reliability coefficient of .70 would note that (a) 70% of the observed score variance is attributable to true score variance and (b) the correlation between the true scores and observed scores is .837 (i.e., $\sqrt{.7}$) for the underlying sample (Crocker & Algina, 1986).

*Non-reporting of reliability coefficients.* Unfortunately, relatively few researchers report reliability coefficients for data from their samples (Onwuegbuzie, 1999; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). For example, Vacha-Haase et al. (1999), who reviewed practices regarding the reporting of reliability coefficients in three journals from 1990 to 1997, found that 64.4% of articles did not provide reliability coefficients for the data being analyzed. Similarly, Vacha-Haase (1998), who identified 628 articles in which the Bem Sex Role Inventory (Bem, 1981) was utilized, found that 86.9% of the articles did not present any score reliability information for the underlying data. Similarly, Simmelink and Vacha-Haase (1999) reported that 75.9% fell into this category with respect to the use of the Rosenberg Self-Esteem Instrument (Rosenberg, 1965). Finally, in a review of 36 research articles published in the 1998 volume of the *British Journal of Educational Psychology*, Onwuegbuzie (1999) found that 72.2% of studies did not report reliability indices for their own sample.

The trend of not reporting current-sample reliability coefficients stems, in part, from a failure to realize that reliability is a function of scores, not of instruments (Thompson & Vacha-Haase, 2000). The dearth in the reporting of reliability estimates led the American Psychological Association (APA) Task Force on Statistical Inference recently to recommend that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (Wilkinson & Task Force on Statistical Inference, 1999, p. 21).

Without information about score reliability, it is impossible to assess accurately the extent to which statistical power is affected. Thus, reliability coefficients always should be reported for the underlying data. Moreover, the use of confidence intervals around reliability coefficients is advocated, considering that reliability coefficients represent only point estimates. The coefficient alpha and KR-20 estimates outlined above both stem from the assumption that each item on an instrument represents a perfectly parallel subscale. Unfortunately, this assumption is extremely tenuous in most situations. As such, both measures of internal consistency yield estimates that represent a lower bound of the theoretical reliability estimate (Crocker & Algina, 1986). Thus, in addition to reporting reliability coefficients for their own sample, researchers should report one-sided (i.e., upper-tailed) confidence intervals for these estimates.

Because internal consistency estimates are essentially a type of correlation coefficient, as are all other estimates of reliability, the sampling distribution of the sample reliability coefficient for all values of the theoretical reliability coefficient other than 0 is skewed. Therefore, the reliability coefficient ($r_{x \cdot x}$) must be transformed in such a way that it has a sampling distribution that is approximately normal. An appropriate transformation is *Fisher's* $Z$ transformation. This transformation statistic is defined as

$$|Z| = 0.5 \log_e \left( \frac{1 + |r_{xx}|}{1 - |r_{xx}|} \right)$$

where $\log_e$ is the natural logarithm and the "| |" indicates that the number contained in it can be either positive or negative. Alternatively, one can obtain Fisher Z-values from tables that are provided in many standard statistics textbooks. Such tables give the value of $Z$ for values of r from 0 to 1.00. (If $r_{x \cdot x}$ is negative, the $Z$ value obtained becomes negative). If the exact value of $r_{x \cdot x}$ is not listed, interpolation is used to obtain the corresponding $Z$-value. Conveniently, the distribution of $Z$ is approximately normal regardless of the size of $n$, with a mean $Z_\rho$, which corresponds to $\rho_{x \cdot x}$ (the theoretical reliability), and a standard deviation given by

$$\sigma_z = \frac{1}{\sqrt{n - 3}}$$

A (1 - $\alpha$)% upper-tailed confidence limit for $Z_\rho$ is

$$Z + (Z_{\alpha}) \cdot \sigma_{z}$$

Thus, the procedure for constructing an upper 95% confidence limit for a reliability coefficient is as follows:

1. Transform the reliability coefficient to Fisher *Z* using the equation above or the Fisher's *Z* transformation table.
2. Compute the standard error of *Z*.
3. Find a (1 - α)% upper confidence limit for $Z_{\rho}$
4. Use the Fisher's *Z* table to transform the upper confidence limits for $Z_{\rho}$ back to the reliability coefficient value.

For example, for a sample of 30 individuals, a reliability alpha reliability coefficient of .80 would yield a 95% upper confidence limit of .89. This interval indicates that over repeated administrations of the same instrument, we expect the true reliability coefficient to lie between .80 (i.e., $r_{x\cdot x}$) and .89 approximately 95% of the time. For a sample of size 50, the same reliability coefficient will yield a 95% upper confidence limit of .87; for a sample of 100 individuals, the upper limit would be .86; and for a sample of size 1000, the corresponding 95% upper confidence limit would be .82. Thus, as with all confidence intervals, for a given reliability coefficient and confidence level, the upper confidence limit decreases as the sample size increases.

Not only is an upper confidence limit likely to provide more accurate information about the theoretical reliability coefficient than is the point estimate (i.e., $r_{xx}$) alone, but this limit could also be used to determine the extent to which a current-sample reliability coefficient differs from the inducted-sample reliability coefficient (i.e., the reliability coefficient reported by the instrument developers). Specifically, if the inducted-sample reliability coefficient is captured by the upper limit (i.e., lies between the current-sample reliability coefficient and its upper limit), it can be inferred that the scores from the current sample generate a reliability coefficient that is similar to that of the inducted sample. Conversely, if the inducted-sample reliability coefficient is greater than the upper limit, the researcher should conduct follow-up analyses to determine whether the lower reliability generated by the present sample reflects any differences in sample composition (Vacha-Haase et al., in press) or homogeneity of score variance (Roberts & Onwuegbuzie, 2000). Information obtained from this follow-up analysis should allow the investigator to put findings into a more appropriate context.

When current-sample reliability coefficients are not available (as is the case when archival data are utilized), researchers, at the very least, should compare the sample composition and variability of scores of the present sample with those of the inducted (i.e., norm) group (Vacha-Haase et al., in press). However, this is not a substitute for obtaining a reliability estimate for the sample (had the researcher been able to do so). Hence, assuming that previously-reported reliability coefficients generalize to a given sample is only marginally justifiable even if the compositions and the score variabilities of the two samples are similar.

In such cases, Magnusson's (1967) formula can be used to approximate the reliability of the present sample, based on the reliability of the inducted sample and the standard deviations of the inducted and current samples, as follows:

$$R_c = 1 - \frac{\sigma_i^2(1 - R_i)}{\sigma_c^2}$$

where $R_c$ = the predicted reliability of the current sample, $R_i$ = the predicted reliability of the inducted sample, $\sigma_c^2$ = the variance of the total instrument scores for the current sample, and $\sigma_i^2$ = the variance of the total instrument scores for the inducted sample. However, it should be noted that these predicted reliabilities are purely theoretical. (For an example of the use of this formula see Diamond and Onwuegbuzie, in press.)

*Conducting tests of statistical significance for reliability coefficients*. Of the relatively few researchers who report reliability coefficients for their sample, some fall into the unfortunate habit of testing these coefficients for statistical significance using the nil null hypothesis (Huck, 2000). However, as Thompson (e.g., Thompson, 1994, 1996, 1998, 1999) and Daniel and his colleagues (Daniel, 1998; Onwuegbuzie & Daniel, in press; Witta & Daniel, 1998) have argued, such tests are inappropriate, because large reliability coefficients typically are statistically significant even when the sample sizes that underlie them are small. In fact, small reliability coefficients will eventually become statistically significant as the sample size increases (Huck, 2000), due to the influence of sample size on statistical significance tests (see for example, Onwuegbuzie & Daniel, 2000). Additionally, because reliability coefficients are sample specific, statistically significant coefficients are neither necessarily replicable not generalizable (Witta & Daniel, 1998). Therefore, rather than utilizing statistical significance tests of reliability coefficients, researchers should assess the (effect) size of reliability estimates to determine the adequacy of instrument scores generated with specific samples.

For example, Nunnally and Bernstein's (1994) criteria could be used for assessing the reliability of scores on non-cognitive measures for a specific sample. According to Nunnally and Bernstein, reliability coefficients of .70 and above should be considered adequate. For scores on measures of cognitive performance, .80 could be utilized as the “cut-off” criterion (e.g., Sattler, 1990), or the more stringent cut-off of .90 could be used (e.g., Gay & Airasian, 2000). Reliability coefficients might be considered adequate even if somewhat lower than any of these criteria depending on how much error the researcher is willing to tolerate in a given study (Pedhazur & Schmelkin, 1991). Further, rather than comparing the sample-specific reliability coefficient to these criteria, the one-sided confidence interval could be used.

*Incorrect reporting of reliability of difference scores.* Researchers often are interested in determining the effect of an intervention by comparing scores on the same instrument administered both before and after the intervention phase. The few investigators who report reliability coefficients for their sample tend to report only the pre-intervention estimates or the post-intervention estimates. However, it is more appropriate to estimate the reliability of difference scores (Allen & Yen, 1979). As noted by Crocker and Algina (1986), the formula for the reliability of the difference between two scores on the same instrument is given by the following:

$$\frac{r_{xx}\sigma_x^2 + r_{yy}\sigma_y^2 - 2r_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y}$$

where $r_{x \cdot x}$ is the pre-intervention reliability estimate, $r_{y \cdot y}$ is the post-intervention reliability estimate, $r_{x \cdot y}$ is the correlation between the pre- and post-intervention scores, $\sigma_x^2$ is the variance of the pre-intervention scores, and $\sigma_y^2$ is the variance of the post-intervention scores. Interestingly, this same formula can be used to determine the reliability of difference scores when scores generated from two instruments measured on the same sample are being compared.

*Assuming Invariance of reliability estimates across sub-samples*. As noted above, although the APA Task Force has advocated that researchers report current-sample reliability coefficients, these recommendations do not go far enough. We contend that reliability coefficients should not only be reported for the full sample at hand, but also for sample subgroups. By reporting only full-sample reliability coefficients, researchers assume that the score reliability is invariant across sub-samples. Yet, closer examination of the coefficient alpha (or KR-20) for most any given dataset will indicate that reliability invariance should not be assumed. For instance, for the two-sample case, scores for both samples would generate identical reliability coefficients if and only if the coefficient alphas are identical, that is, if and only if

$$\frac{k_1}{k_1 - 1}\left(1 - \frac{\Sigma\hat{\sigma}_{1i}^2}{\Sigma\hat{\sigma}_{1x}^2}\right) = \frac{k_2}{k_2 - 1}\left(1 - \frac{\Sigma\hat{\sigma}_{2i}^2}{\Sigma\hat{\sigma}_{2x}^2}\right)$$

where the equation on the left hand side represents the coefficient alpha for the first sample, and the equation on the right hand side represents the coefficient alpha for the second sample. Because both samples would have been administered the same instrument, $k_1 = k_2$, the equation above reduces to the following:

$$\left(1 - \frac{\Sigma\hat{\sigma}^2_{1i}}{\Sigma\hat{\sigma}^2_{1x}}\right) = \left(1 - \frac{\Sigma\hat{\sigma}^2_{2i}}{\Sigma\hat{\sigma}^2_{2x}}\right)$$

which further implies that when the reliability estimates are equal,

$$\left(\frac{\Sigma\hat{\sigma}^2_{1i}}{\Sigma\hat{\sigma}^2_{1x}}\right) = \left(\frac{\Sigma\hat{\sigma}^2_{2i}}{\Sigma\hat{\sigma}^2_{2x}}\right)$$

Thus, for the reliability coefficients of two sub-samples to be equal, it is not enough for the variance of the total instrument scores to be equal. The sum of the individual item variances must also be equal for both sub-samples (or the ratio of the sum of the individual item variances to the variance of the total instrument scores must be equal for both sub-samples). Bearing in mind the fluctuations that occur in item responses, it is tenuous to assume that the sum of the individual item variances would be equivalent from one sub-sample to the next. Moreover, the interest of the researcher in comparing subgroups usually stems from an expectation that these subgroups are different with respect to the dependent variable(s). Thus, under this assumption, it seems counter-intuitive to expect the variances of the total instrument scores across sub-samples to be equal. It is for this reason that, when conducting an independent *t*-test, it is typically recommended that equal variances not be assumed (Onwuegbuzie & Daniel, 2000). Further, it is even less plausible to expect the sum of the individual item variances to be equal. Thus, even if the total instrument scores of both subgroups are identical, it should not be assumed that the reliability estimates also will be equal because, as noted above, the sum of the individual item variances also must be equivalent.

Disturbingly, it is possible for a reliability coefficient from the full sample not only to mask notable differences in sub-sample reliabilities, but also to conceal low score reliabilities generated from scores of one or more of the subgroups. For example, in a two-sample case, it is feasible to obtain a large reliability estimate for the full sample even when the reliability coefficient of one group is relatively large but the coefficient for the other group is relatively small. In fact, it is likely that such a case would produce a different outcome in terms of statistical and practical significance than would a scenario in which the ratio of reliability coefficients is much smaller. Simply put, comparing subgroups with different reliability coefficients can affect Type I and Type II error rates, as well as effect size estimates. The following examples using two small heuristic datasets will serve to illustrate the adverse effects of differential reliability on statistical power.

## Heuristic Examples

For the purposes of the current discussion, two heuristic datasets were utilized that were (hypothetically) generated from the same 6-item instrument. Both datasets contained dichotomous (i.e., 0/1) scores from two groups, each containing 64 cases. This sample size ($n$ = 128) was selected via an *a priori* power analysis because it provided acceptable statistical power (i.e., .80) for detecting a moderate difference in means (i.e., Cohen's [1988] $d$ = .5) at the (two-tailed) .05 level of statistical significance (Erdfelder, Faul, & Buchner, 1996).

The first dataset (Table 1), hereafter termed the *invariant-reliability dataset*, was designed such that scores of both subgroups yielded adequate classical theory alpha reliability coefficients; that is, both coefficients were greater than .7, using Nunnally and Bernstein's (1994) criteria, as was the full-sample reliability estimate. On the other hand, the second dataset (Table 2), hereafter termed the *variant-reliability dataset*, was constructed such that although scores from the full sample yielded an adequate reliability coefficient, only the first group generated an adequate reliability estimate (the same data were used for Group 1 as in the invariant-reliability dataset), whereas scores from the second subgroup yielded a low reliability coefficient.

Insert Tables 1 and 2 about here

The summary statistics for the invariant-reliability dataset are presented in Table 3. It can be seen from this table that whether or not equal variances are assumed (we strongly advocate that equal variances should *not* be assumed), the difference between the group means is statistically significant. Conversely, even though the respective group means in the variant-reliability dataset are identical to those in the invariant-reliability dataset, the difference in means is no longer statistically significant. Thus, the lower reliability pertaining to Group 2 in the variant-reliability dataset is associated with relatively low statistical power. In fact, this dataset led to the opposite conclusion with respect to statistical significance. This example, thus, provides empirical evidence that statistical power is affected by low subgroup reliability.

Interestingly, other variant-reliability datasets (not presented) were constructed that also led to statistical nonsignificance. These findings suggested that (a) researchers should not assume that subgroup reliabilities are equal, and (b) researchers should report subgroup reliability coefficients whenever possible, alongside their confidence intervals.

Insert Tables 3 and 4 about here

## Discussion

The major purpose of the present paper was to identify errors made by researchers when dealing with reliability coefficients and to outline best practices for these indices. Common errors made by researchers include (a) incorrectly stating that instruments are reliable, (b) incorrectly interpreting correlation coefficients, (c) not reporting reliability coefficients for their own sample, (d) conducting tests of statistical significance on reliability coefficients, and (e) failing to report reliability of difference scores. Additionally, several recommendations were made. First and foremost, consistent with the recommendations of the APA task force (Wilkinson & the Task Force on Statistical Inference, 1999) and others (e.g., Onwuegbuzie, 1999; Thompson & Vacha-Haase, 2000; Vacha-Haase et al., 1999), it is advocated that researchers report reliability coefficients for their underlying data.

Second, when sample-specific reliability coefficients are not available (as is the case when archival data are utilized), it is recommended that researchers not rely solely on reporting reliability indices provided by instrument developers (i.e., from the inducted sample), but should compare the sample composition and variability of scores of the present sample with those of the inducted (i.e., norm) group (Vacha-Haase et al., in press). For situations when the current-sample reliability estimates are not obtainable, it is recommended that researchers approximate the score reliability for their sample using Magnusson's (1967) formula.

In addition to the above recommendations, we offer two additional recommendations pertaining to (a) use of confidence intervals when reporting and interpreting reliability coefficients and (b) routine reporting and interpretation of sub-sample reliabilities in group comparison studies. With respect to the former, because reliability coefficients represent point estimates that are subject to error, and because the reliability estimates represent a lower bound for the theoretical reliability coefficient, researchers should provide upper *Z*-transformed confidence limits alongside reliability estimates. Surprisingly, our review of the literature did not yield any recommendations for such confidence intervals to be presented, and no mention of confidence intervals for this purpose was made in measurement textbooks we consulted (e.g., Allen & Yen, 1979; Crocker & Algina, 1976; Magnusson, 1967; Mehrens & Lehmann, 1991). Nevertheless, these confidence intervals around the reliability coefficient can be compared to coefficients presented in instrument manuals to assess generalizability.

Unfortunately, computing upper confidence limits for reliability coefficients will be cumbersome for most researchers, clinicians, and practitioners. This is because algorithms or statistical tables are needed to obtain Fisher *Z*-values. Thus, it is recommended that creators of the major statistical packages consider providing upper confidence limits for reliability estimates.

Finally, as previously noted, the excellent recommendations of the American Psychological Association Task Force regarding the reporting of current-sample reliability coefficients do not go far enough. Reliability coefficients should be reported not only for the full sample at hand, but also for each subgroup. Our heuristic examples for a two-sample case (i.e., *t*-test) have illustrated how comparing subgroups with different reliability coefficients can affect statistical power. Obviously, simulation studies are needed to examine further the extent to which variant-reliability sub-samples affect Type I and Type II error rates, as well as effect size estimates. Nevertheless, the heuristic examples presented herein, along with other sub-samples examined but not presented herein, suggest that sub-samples with scores that generate markedly different reliability estimates are problematic, even when the full-sample reliability coefficients are adequate.

Authors of statistics textbooks routinely report that statistical power is affected by at least three components: (a) sample size, (b) level of statistical significance, and (c) effect size. However, as shown in the heuristic examples provided above, a fourth component should be added, namely, the reliability of scores. Disturbingly, in examining other variant-reliability datasets, we found that it is possible for scores of one subgroup to yield a *negative* reliability coefficient and when the full-sample reliability coefficient is positive and of adequate magnitude (i.e., greater than .70).

## References

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

American Educational Research Association, American Psychological Association, & National Council on Measurement and Evaluation. (1985). *Standards for educational and psychological testing.* Washington: American Psychological Association.

Bem, S.L. (1981). *Bem Sex-Role Inventory: Professional manual.* Palo Alto, CA: Consulting Psychologists Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: John Wiley.

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

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

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for editorial policies of educational journals. *Research in the Schools, 5*, 23-32.

Diamond, P.J., & Onwuegbuzie, A.J. (in press). Factors associated with reading achievement and attitudes among elementary school-aged students. *Research in the Schools.*

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1-11.

Gay, L.R., & Airasian, P.W. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Englewood Cliffs, N.J.: Prentice Hall.

Huck, S.W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.

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

Magnusson, D. (1967). *Test theory*. Boston, MA: Addison-Wesley.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth, TX: Holt, Rinehart, and Winston.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20.

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

Onwuegbuzie, A.J. (1999, September). *Common analytical and interpretational errors in educational research*. Paper presented at the annual meeting of the European Conference on Educational Research (ECER), Lahti, Finland.

Onwuegbuzie, A.J., & Daniel, L.G. (2000, April). *Common analytical and interpretational errors in educational research*. Paper presented at the annual conference of the American Educational Research Association (AERA), New Orleans.

Onwuegbuzie, A.J., & Daniel, L.G. (in press). Uses and misuses of the correlation coefficient. *Research in the Schools*.

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

Roberts, J.K., & Onwuegbuzie, A.J. (2000, November). Alternative approaches for interpreting alpha with homogeneous sub-samples. In G. Halpin (Chair), *Reliability Issues*. Symposium to be conducted at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sattler, J.M. (1990). *Assessment of children*. San Diego, CA: Author.

Simmelink, S., & Vacha-Haase, T. (1999). *Reliability generalization with the Rosenberg Self-Esteem Instrument*. Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Fort Collins, CO.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

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*(2), 26-30.

Thompson, B. (1998, April). *Five methodological errors in educational research: The pantheon of statistical significance and other faux pas*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Thompson, B. (1999, April). *Common methodological mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. Retrieved October 11, 1999 from the World Wide Web: **http://acs.tamu.edu/~bbt6147/aeraad99.htm.**

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

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., Kogan, L. R., & Thompson, B. (in press). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*(4).

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

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

Witta, E.L., & Daniel, L.G. (1998, April). *The reliability and validity of test scores: Are editorial policy changes reflected in journal articles?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

*Table 1*

*Invariant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

*Table 1 (Cont/d...)*

*Invariant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |

*Table 1 (Cont/d...)*

*Invariant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |

*Table 1 (Cont/d...)*

*Invariant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Table 2*

*Variant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

*Table 2* (Cont/d...)

*Variant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 4 |
| 2 | 1 | 0 | 0 | 1 | 0 | 1 | 3 |
| 2 | 1 | 1 | 0 | 0 | 1 | 1 | 4 |
| 2 | 1 | 0 | 1 | 0 | 0 | 1 | 3 |
| 2 | 1 | 1 | 0 | 1 | 0 | 0 | 3 |

*Table 2* (Cont/d...)

*Variant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| 2 | 1 | 0 | 1 | 1 | 0 | 1 | 4 |
| 2 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 2 | 0 | 0 | 1 | 1 | 0 | 0 | 2 |
| 2 | 1 | 1 | 1 | 1 | 0 | 0 | 4 |
| 2 | 1 | 0 | 1 | 0 | 1 | 0 | 3 |
| 2 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 | 2 |
| 2 | 1 | 0 | 0 | 0 | 1 | 0 | 2 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 | 4 |
| 2 | 1 | 0 | 1 | 0 | 1 | 0 | 3 |
| 2 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 | 2 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 1 | 0 | 0 | 2 |
| 2 | 0 | 0 | 0 | 1 | 0 | 1 | 2 |
| 2 | 0 | 0 | 1 | 1 | 0 | 1 | 3 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 2 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 0 | 1 | 1 | 0 | 0 | 1 | 3 |
| 2 | 0 | 1 | 1 | 0 | 0 | 1 | 3 |

*Table 2* (Cont/d...)

*Variant-Reliability Dataset*

| Group | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Total |
|---|---|---|---|---|---|---|---|
| 2 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 | 2 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 2 | 0 | 1 | 1 | 0 | 0 | 0 | 2 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 1 | 0 | 0 | 2 |
| 2 | 0 | 1 | 0 | 1 | 1 | 0 | 3 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Table 3*

*Summary Statistics for Invariant-Reliability Dataset*

|  |  |  |  | Equal Variances |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  |  |  | Assumed |  | Not Assumed |  |
| Index | Full Sample | Group 1 ($n$ = 64) | Group 2 ($n$ = 64) | *t*-value | *df* | *t*-value | *df* |
| Reliability | .83 | .89 | .71 |  |  |  |  |
| 95% Upper Confidence Limit | .87 | .93 | .80 |  |  |  |  |
| Mean | 2.73 | 3.05 | 2.42 | 1.99* | 126 | 1.99* | 101 |
| Standard Deviation | 1.79 | 2.17 | 1.26 |  |  |  |  |

* p < .05

*Table 4*

*Summary Statistics for Variant-Reliability Dataset*

| | | | | Equal Variances<br>Assumed | | Equal Variances<br>Not Assumed | |
|---|---|---|---|---|---|---|---|
| Index | Full Sample | Group 1 ($n$ = 64) | Group 2 ($n$ = 64) | $t$-value | $df$ | $t$-value | $df$ |
| Reliability | .79 | .89 | .66 | | | | |
| 95% Upper Confidence Limit | .84 | .93 | .76 | | | | |
| Mean | 2.73 | 3.05 | 2.42 | 1.78 | 126 | 1.78 | 121.34 |
| Standard Deviation | 2.00 | 2.17 | 1.78 | | | | |

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: Reliability Generalization: The Importance of Considering Sample Specificity, Confidence Intervals, and Subgroup Differences

Author(s): Anthony J. ONWUEGBUZIE + LARRY G. DANIEL

Corporate Source:

Publication Date: 2000

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

*I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.*

Sign here, → please

Signature:

Organization/Address:

Anthony J. Onwuegbuzie, Ph.D. F.S.S
Department of Educational Leadership
College of Education
Valdosta State University
Valdosta, Georgia 31698

Printed Name/Position/Title: ANTHONY J. ONWUEGBUZIE, ASST PROFESSOR

Telephone: 912-247-8333

FAX: 912-247-8326

E-Mail Address: TONWUEGB@VALDOSTA.EDU

Date: 11/23/00

(over)