ED 464 100 TM 033 781

AUTHOR Onwuegbuzie, Anthony J.; Daniel, Larry G.; Roberts, J. Kyle
TITLE A Proposed New "What If Reliability" Analysis for Assessing the Statistical Significance of Bivariate Relationships.
PUB DATE 2001-11-00
NOTE 31p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (30th, Little Rock, AR, November 14-16, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Correlation; *Reliability; Sample Size; *Statistical Significance

ABSTRACT

The purpose of this paper is to illustrate how displaying disattenuated correlation coefficients along with their unadjusted counterparts will allow the reader to assess the impact of unreliability on each bivariate relationship. The paper also demonstrates how a proposed new "what if reliability" analysis can complement the conventional null hypothesis significance test of bivariate relationships. The "what if reliability" procedure helps the researcher determine the extent to which the sample size, as opposed to the effect size, is responsible for the observed statistically significant, or not statistically significant, finding. The analyses illustrate how the sample size needed to detect a statistically significant bivariate relationship decreases as the observed score reliability coefficient pertaining to the independent or dependent measure (theoretically) increases, holding all other factors constant. As such, "what if reliability" analyses will help researchers interpret their results by considering the extent to which the reliability coefficients contribute or fail to contribute to the ability to achieve a statistically significant result. (Contains 3 tables and 28 references.) (SLD)

Running head: INDICES OF SCORE RELIABILITY

A Proposed New "What if Reliability" Analysis for Assessing the Statistical Significance of Bivariate Relationships

Anthony J. Onwuegbuzie

Howard University

Larry G. Daniel

University of North Florida

J. Kyle Roberts

University of North Texas

Correspondence should be addressed to Anthony J. Onwuegbuzie, Department of Human Development and Psychoeducational Studies, School of Education, Howard University, 2441 Fourth Street, NW, Washington, DC 20059, or E-Mail: (tonyonwuegbuzie@aol.com).

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 14-16, 2001.

Abstract

The purpose of the present paper is twofold. First, the authors illustrate how displaying disattenuated correlation coefficients alongside their unadjusted counterparts will allow the reader to assess the impact of unreliability on each bivariate relationship. Second, the authors demonstrate how a proposed new "what if reliability" analysis can complement the conventional null hypothesis significance test (NHST) of bivariate relationships. Such analyses illustrate how the sample size needed to detect a statistically significant bivariate relationship decreases as the observed score reliability coefficient pertaining to the independent and/or dependent measure (theoretically) increases, holding all other factors constant. As such, "what if reliability" analyses will help researchers to interpret their results by considering the extent to which the reliability coefficient(s) contributed, or failed to contribute to the ability to achieve a statistically significant result.

A Proposed New "What if Reliability" Analysis

for Assessing the Statistical Significance of Bivariate Relationships

One of the assumptions underlying null hypothesis significance tests (NHST) is that all variables involved are measured without error (Myers, 1986; Onwuegbuzie & Daniel, in press-a). Unfortunately, when measurement errors are present, as is typically the case in the social and behavioral sciences, the relationships computed from the sample data will systematically underestimate the strength of the associations in the population. Indeed, indices of both statistical significance and practical significance will be adversely affected by errors in measurement (Onwuegbuzie, 2001).

In the two-variable case, errors of measurement yield biased estimates of correlation coefficients that attenuate the true relationships. In fact, the greater the measurement error, the more the correlation coefficient is attenuated. Thus, knowledge of the error of estimate is vital. A common way of assessing reliability is via the reliability coefficient. Indeed, the relationship between the reliability and the standard error of measurement can be seen from the following formula:

$$\sigma_{\Sigma} = \sigma_{\acute{\varepsilon}} \sqrt{1 - r_{\acute{\varepsilon}\acute{\varepsilon}}}$$

where σ_E is the standard error of measurement, σ_x is standard deviation of the full scale scores, and r_{xx} = is the reliability yielded by the scale scores. Thus, the lower the score reliability, the larger the standard error of measurement. As such, the reliability estimate provides valuable information about the measurement error. Unfortunately, most researchers do not report reliability estimates for their own data (Henson, 2000;

Onwuegbuzie & Daniel, 2000, 2001, in press-b; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998). In fact, in studies examining whether researchers report reliability coefficients, the proportion of researchers who do not report reliability coefficients for data from their underlying sample has been found to range from 64.4% (Vacha-Haase, Ness, Nillsson, & Reetz, 1999) to 86.9% (Vacha-Haase, 1998). Hence, many researchers are not in a position to determine the extent to which measurement error affected the observed findings in their study.

In an attempt to increase substantially the proportion of researchers who report sample-specific reliability estimates, Pedhazur and Schmelkin (1991) asserted that

> Researchers who bother at all to report reliability estimates for [scores on] the instruments they use frequently report only reliability estimates contained in the manuals of the instruments or estimates reported by other researchers. Such information may be useful for comparative purposes, but it is imperative to recognize that the relevant reliability estimate is the one obtained for the sample used in the [current] study under consideration. (p. 86)

More recently, the American Psychological Association (APA) Board of Scientific Affairs, who convened a committee called the Task Force on Statistical Inference for the purpose of providing recommendations for the use of statistical methods offered the following suggestion for researchers:

> If a questionnaire is used to collect data, summarize the psychometric properties of its scores with specific regard to the way the instrument is used in a population. Psychometric properties include measures of validity, reliability, and any other

qualities affecting conclusions....Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of the scores. (Wilkinson & the Task Force on Statistical Inference, 1999, p. 5)

The latest version of the American Psychological Association (APA), version 5, provided some much-needed recommendations and stipulations for presenting "informationally adequate statistics" (APA, 2001, p. 23), including the reporting of effect sizes and confidence intervals. Clearly, the APA Task Force played an important role here. Unfortunately, despite the recommendations of the Task Force noted above, only one suggestion was given by APA (2001) regarding the delineation of sample-specific reliability estimates:

For correlational analyses (e.g., multiple regression analysis, factor analysis, and structural equation modeling), the sample size and variance-covariance (or correlation) matrix are needed, accompanied by other information specific to the procedure used (e.g., variable means, **reliabilities**, hypothesized structural models, and other parameters)..." [emphasis and end parenthesis added] (p. 23)

Morever, the fact that the word *reliabilities* was included only in parentheses gives the impression that the reporting of sample-specific reliability coefficients is not of primary importance, or even optional. Thus, it is likely that the majority of researchers will continue to fail to report these indices, let alone to report them consistently. Nor is rhetoric provided by a handful of research methodologists sufficient to reverse this trend.

Rather, what is needed is more compelling evidence of how information about current-sample specific reliability estimates can facilitate data analysis and interpretation.

Therefore, the objective of the present paper is to provide researchers with further incentive for reporting reliability coefficients for their own data. Specifically, the purpose of the current paper is twofold. First, the authors illustrate how displaying disattenuated correlation coefficients alongside their unadjusted counterparts will allow the reader to assess the impact of unreliability on each bivariate relationship. Second, the authors demonstrate how a proposed new "what if reliability" analysis can complement the conventional NHST of bivariate relationships. Such analyses indicate how the sample size needed to detect a statistically significant bivariate relationship decreases as the observed score reliability coefficient pertaining to the independent and/or dependent measure (theoretically) increases, holding all other factors constant. As such, it is contended that "what if reliability" analyses will help researchers to interpret their results by considering the extent to which the reliability of scores on one or both variables affect the statistical significance of the bivariate correlation coefficient.

Correction for Attenuation

In every empirical study, researchers strive to utilize measures that yield scores that are as error-free as possible. While this goal often is met for empirical research conducted in the physical and life sciences, this is seldom the case when dealing with social, behavioral, psychological, sociological, and educational constructs. Unfortunately, when measurement errors prevail, any relationship that is derived from the underlying data will systematically underestimate the strength of the association in the population

(Huck, 2000). In other words, errors of measurement produce negatively biased estimates of the observed relationship that attenuate the true relationship. As noted earlier, the greater the measurement error, the more the association is attenuated.

Researchers and authors of statistics textbooks alike have routinely acknowledged that statistical power is affected by the following three components: (a) size of the sample, (b) level of statistical significance (i.e., Type I error probability allowed for), and (c) effect size. However, as admonished by Onwuegbuzie and Daniel (2000), the role of score reliability is often neglected. Even APA (2001) failed to mention the influence of score reliability on statistical power, as illustrated in the following statement:

> Take seriously the statistical power consideration associated with your tests of hypotheses. Such considerations relate to the likelihood of correctly rejecting the tested hypotheses, given a particular alpha level, effect size, and sample size. In that regard, you should routinely provide evidence that your study has sufficient power to detect effects of substantive interest (see Cohen, 1988). You should be similarly aware of the role played by sample size in cases in which not rejecting the null hypothesis is desirable (i.e., when you wish to argue that there are no differences)...(p. 25)

The way in which score reliability can affect statistical power recently was illustrated by Onwuegbuzie and Daniel (2000), who demonstrated that subgroups with scores that generate markedly different reliability estimates can seriously reduce statistical power, even when the full-sample reliability coefficients are adequate. Specifically, Onwuegbuzie and Daniel produced two datasets that consisted of scores pertaining to two sub-samples. The first dataset, which they called the *invariant-reliability dataset* (p. 23), was constructed such that scores of both subgroups yielded adequate classical theory alpha reliability coefficients; namely, .89, and .71 (Nunnally & Bernstein, 1994); also, the full-sample reliability estimate of .83 was adequate. Conversely, the second dataset, which the authors termed the *variant-reliability dataset* (p. 23), was generated such that although scores from the full sample yielded an adequate reliability coefficient (i.e., .79), only the first subgroup generated an adequate reliability estimate (i.e., .89), whereas scores from the second subgroup yielded a low reliability coefficient (i.e., .66). Onwuegbuzie and Daniel found that the invariant-reliability dataset, containing adequate subgroup reliability estimates yielded a statistically significant difference between the two groups. On the other hand, the variant-reliability dataset yielded no statistically significant difference, even though the respective group means in the variant-reliability dataset were identical to those in the invariant-reliability dataset.

Onwuegbuzie and Daniel (2000) recommended that researchers not only report reliability coefficients for the full sample at hand, but also for each subgroup. Indeed, providing sub-sample reliability estimates alongside those for the full sample is consistent with APA's (2001) following recommendations: (a) "report the data in sufficient detail to justify the conclusions" (p. 20); and (b) "When reporting inferential statistics, include sufficient information to help the reader fully understand the analyses conducted and possible explanations for the outcomes of these analyses" (p. 23).

Because low reliability coefficients tend to reduce statistical power, in cases when the null hypothesis is not rejected, and one or more of the measures generate scores that have a low reliability coefficient, the researcher cannot be certain whether the statistically nonsignificant finding represents a real outcome or a statistical artifact. This is true in both the bivariate case and the multivariate case. With respect to the former, some theorists recommend use of sample-specific reliability coefficients to adjust correlation coefficients to account for the estimated amount of unreliability. These analysts use a correction-for-attenuation formula which yields an adjusted/disattenuated correlation coefficient that is *always* higher than the uncorrected, raw *r* (Huck, 2000), with the exception of the utopian case of both variables being measured with absolutely no error.

Three correction-for-attenuation formulae are commonly used. Some theorists utilize Spearman's (1910) double correction formula:

$$\rho_{\cdot\cdot} = \frac{r_{\cdot\cdot}}{\sqrt{r_{\cdot\cdot}}\sqrt{r_{\cdot\cdot}}}$$

where ρ_{xy} is the corrected correlation coefficient, r_{xy} is the obtained sample correlation coefficient, r_{xx} is the reliability of scores yielded by the measure of the independent variable, and r_{yy} is the reliability of scores generated by the measure of the dependent variable. This formula corrects for unreliability of scores generated by measures of both the independent and dependent variables. This is how the formula works. Suppose the correlation coefficient between an independent variable and a dependent variable was .30, representing a moderate relationship (Cohen, 1988). Let us also suppose that the dependent measure yielded scores that produced a reliability estimate of .80, whereas the independent measure generated scores that resulted in a reliability coefficient of .70.

Using these values, the double correction formula would lead to a corrected correlation coefficient of

$$\rho_{..} = \frac{.30}{\sqrt{.70}\,\sqrt{.80}} = 0.40$$

The adjusted coefficient indicates that if scores pertaining to the independent and dependent measures had generated perfect reliability (i.e., no measurement error), holding all other data constant, the correlation coefficient would have been .40, which is 33.3% larger. This illustrates how unreliability in both measures adversely affects statistical power and effect sizes of bivariate relationships.

Other measurement theorists advocate the single correction formula

$$\rho_{..} = \frac{r_{..}}{\sqrt{r_{..}}}$$

or

$$\rho_{..} = \frac{r_{..}}{\sqrt{r_{..}}}$$

where the first single correction formula corrects for unreliability in scores generated by measures of the dependent variable only (i.e., *dependent variable-based corrected correlation coefficient*), and the second single correction formula adjusts for unreliability in scores yielded by measures of the independent variable only (i.e., *independent variable-based corrected correlation coefficient*). These single correction formulae are useful in cases in which the reliability of scores on one of the variables is unknown to the researcher (e.g., when using data from standardized achievement tests in which the raw scores are not reported by the agency administering the test). If the correlation coefficient was again .3 and the reliability estimate for the dependent variable was .8, with the reliability measure of the independent variable being unknown, then the *dependent variable based corrected correlation coefficient* would be

$$\rho_{\cdot\cdot} = \frac{.30}{\sqrt{.80}} = 0.34$$

This represents a 13.3% increase in the reliability index. Further, if the correlation coefficient was again .3 and the reliability estimate for the independent variable was .7, with the reliability measure of the dependent variable being unknown, then the *independent variable based corrected correlation coefficient* would be

$$\rho_{\cdot\cdot} = \frac{.30}{\sqrt{.70}} = 0.36$$

This represents a 20.0% increase in the reliability index. A comparison of the three corrected correlation coefficients shows that the double correction formula yields a larger correction than do either of the single correction formulae. Indeed, this is always the case except when the independent or dependent variable is measured without any error (i.e., perfect reliability), in which case the double correction formula will be identical to one of the single correction formulae. As such, both single correction formulae are equivalent to the double correction formula, with the unknown reliability estimate set equal to one. Because it is virtually impossible for score reliability to be equal to unity in the social and behavioral sciences, disattenuated correlation coefficients represent an underestimate. Simply put, it is better to utilize the double correction formula than either of the two single correction formulae because the former uses the full reliability information–providing justification for reporting score reliability for *all* measures, whenever possible.

Table 1 presents corrected correlation values for an observed correlation coefficient of .30 across various combinations of score reliability estimates for the independent and dependent measures. Researchers could use this table after computing correlation coefficients. As noted above, if the score reliability estimate for one of the variables is unknown then the researcher should set it equal to 1.00 in Table 1. Indeed, this table could be used to produce a *table of corrected intercorrelations*.

Insert Table 1 about here

It should be noted that although it is theoretically impossible to compute a reliability coefficient that exceeds 1.00, it is empirically feasible using any of the three correction formulae above to compute such a coefficient. For example, if r_{xy} = .6, r_{yy} = .7, and r_{xx} = .5, then the corrected reliability coefficient, ρ_{xy}, will equal 1.01. Although meaningless, Spearman (1910) attributed this anomalous coefficient to sampling error in both variables, whereas Johnson (1944) contended that this occurrence was due to an inadequate sample size. Consistent with Johnson's assertions, Nunnally (1978) reported that such greater-than-unity reliability estimates will be produced when sample sizes are less than 300. Whatever the explanation for greater-than-unity reliability coefficients, all corrected reliability estimates greater than 1.00 should be truncated to unity.

As informative as this technique can be, correcting for attenuation is an extremely controversial technique because it is subject to misapplication and misinterpretation (Muchinsky, 1996; Onwuegbuzie & Daniel, in press-b). First, some researchers incorrectly claim that the method of correcting for attenuation improves the predictive accuracy of measures; yet, this can never occur. That is, applying a correction for attenuation cannot render a measure more predictive than it actually is (Muchinsky, 1996).

Second, a misapplication stems from the practice by many meta-analysts of disattenuating findings from individual studies before aggregating them into a composite score (i.e., effect size measure). According to Muchinsky (1996), this practice is relatively common. Unfortunately, because researchers are not consistent in the procedures that they employ to estimate reliability (i.e., internal consistency, test-retest, and parallel forms), these meta-analysts end up violating a major assumption of classical measurement theory that invalidates the interchangeability of different types of reliability (Cronbach, 1947). Alternatively stated, aggregating indices that have been disattenuated using different measures of reliability seriously distort the validity of the resultant effect size estimates. Moreover, as noted by Onwuegbuzie and Daniel (in press-b), because many researchers presently do not report sample-specific reliability coefficients in their reports, meta-analysts who choose to disattenuate the correlation coefficients of original researchers are left with missing data. Removing studies from the meta analysis for which no reliability estimates are provided or using arbitrary (e.g., .60; Schmidt, Hunter, & Urry, 1976) imputed estimates or published values for missing reliability coefficients leads to biased composite effect sizes even when score reliability is high (Onwuegbuzie & Daniel, in press-b). The finding that the proportion of researchers who do not report reliability coefficients for their underlying sample ranges from 64.4% to 86.9% (Vacha-Haase, 1998; Vacha-Haase, Ness, Nillsson, & Reetz, 1999) suggests that the amount of bias that prevails for meta analyses conducted by those who disattenuate correlation coefficients typically is large.

A third area of contention is which type of reliability to attenuate. Whereas Johnson (1950) recommended that test-retest reliability coefficients be utilized in correction formulas, Guilford (1954) and others advocated the use of coefficients of equivalence. However, the majority of researchers (e.g., Nunnally & Bernstein, 1994) have promoted the use of internal consistency estimates. Indeed, use of internal consistency appears to be justified because it is the most commonly reported index, and thus should be utilized unless temporal instability is considered as a source of error, in which case the coefficient of stability should be employed (Lord & Novick, 1968).

Perhaps the biggest criticism of the use of disattenuated correlation coefficients is that it may mislead the researcher "into believing that a better correlation has been found than that actually evidenced in the available data" (Nunnally, 1978, p. 237). However, this criticism does not necessarily mean that corrected correlations should not be used. Rather, it suggests that researchers should be extremely careful when interpreting them. Strictly speaking, correction for attenuation represents an estimate of how large the correlation would be if the two underlying variables yielded scores that were perfectly reliable--no more, no less. Indeed, it could be argued that the correction for attenuation label is somewhat misleading because this class of formula represents an upper bound for the observed correlation coefficient rather than a correction. As such, corrected correlation coefficients are theoretical rather than actual values. However, as long as researchers bear this mind, disattenuated correlation coefficientss can provide useful information. For example, presenting both the uncorrected and corrected correlation coefficients will allow the reader to assess the impact of unreliability on each bivariate relationship (Onwuegbuzie & Daniel, in press-b). An even more useful application of correction for attenuation formulae are that they allow the researcher to conduct a "what if reliability" analysis. It is to this that we now turn.

Proposed "What If Reliability" Analysis

Thompson (1989a, 1989b) proposed a "what if" method to facilitate the interpretation of null hypothesis significance tests in a sample size context. Specifically, the "what if" procedure helps the researcher to determine the extent to which the sample size, as opposed to the effect size, was responsible for the observed statistically significant or non-statistically significant finding. Moreover, for a given statistically significant finding, this technique allows the analyst to specify how large a sample is needed to obtain statistical significance for an observed finding in cases in which the null hypothesis is not rejected, as well as how small a sample is needed before an observed statistically significant result is no longer statistically significant. Thompson's (1989a, 1989b) method utilizes the uncorrected effect sizes, which are in the metric of the sample. Kieffer and Thompson (1999) criticized use of the "what if" analyses with uncorrected effect sizes because such effect sizes do not take into account the fact "that the amount of sampling error (and therefore the positive bias in the "uncorrected" effect size) will change as sample size itself changes" (Kieffer & Thompson, 1999, p. 11). In their award-winning work, Kiefer and Thompson (1999) subsequently proposed a "what if" analysis that employs the corrected estimate of the population effect size as the metric for examining the influence of sample size on observed *p*-values.

The proposed "what if reliability" analytic method utilizes a similar logic to that of Kieffer and Thompson (1999), except that it focuses on the reliability coefficient as the effect size index. The "what if reliability" analysis utilizes the fact that as the reliability estimate for the scores on the dependent and/or the independent measure decreases, the difference between the uncorrected and corrected correlation coefficient increases, and, subsequently, the sample size needed for statistical significance of the correlation coefficient decreases. For observed correlation coefficients that are statistically significant, the "what if reliability" analysis determines how small the sample size can be before the measurement-error-free correlation is no longer statistically significant, as a function of the score reliability of the independent and dependent measures. For observed correlation coefficients that are *not* statistically significant, the "what if reliability" analysis enables the researcher either (a) to ascertain how small the sample size *theoretically* can be before the measurement-error-free correlation coefficient is no longer statistically significant, as a function of the score reliability of the independent and dependent measures; or (b) in the case where the corrected correlation coefficient is still not statistically significant, to delineate the *theoretical* upper bound for the observed correlation coefficient (i.e., effect size), as well as the sample size needed to produce a statistically significant finding. The sample size needed in (b) would be larger than the original sample size; however, it would be smaller than that needed for the observed correlation coefficient to be statistically significant.

The steps for conducting the "what if reliability" analysis are as follows. First, the observed correlation coefficient must be computed. Second, the score reliability (i.e., internal consistency) estimates pertaining to the independent and/or the dependent variables should be used in the double-correction formula (setting any unknown reliability estimate equal to 1.00) to determine the corrected correlation coefficient. (Confidence intervals can be constructed around the disattenuated correlation coefficient.) Third, if this corrected correlation coefficient is statistically significant, then the sample size should be determined such that any reduction in cases would yield a statistically non-significant result. Conversely, if the corrected correlation coefficient is not statistically significant, then the sample size should be determined at which the corrected correlation coefficient would become statistically significant.

Table 2 illustrates the "what if reliability" analysis for a moderate observed correlation coefficient (i.e., r = .3; Cohen, 1988) and as a function of the reliability estimate of the independent and dependent variables. (Table 3 presents the Excel spreadsheet commands for conducting the what if reliability" analysis.) For example, from Table1, it can be seen that if the score reliability index for both the independent and dependent variables are .8, a correlation coefficient of .30 would yield a corrected correlation of .375. From Table 2, it can be seen that this corrected correlation would be statistically significant with a sample size as small as 28. On the other hand, if the reliability coefficient for both the independent and dependent variables is .6, a correlation coefficient of .30 would yield a corrected correlation index of .375 (Table 1), which, in turn, would be statistically significant with a sample size as small as 16 (Table 2). Comparing these two sets of findings indicates that, holding all other factors constant, a reduction in the score reliability indices of both measures from .8 to .6 (i.e., 25% reduction) results in a smaller sample size (from 28 to 16, or a 42.9% reduction in number of cases) needed. Thus, the more unreliable the measure(s), the larger the correction to the correlation coefficient, and the less expensive the study would be in terms of sample size needed to obtain statistical significance, if the scores for both variables had yielded perfect reliability--again, holding everything else constant.

Insert Table 2 about here

Insert Table 3 about here

Discussion

The illustrative results presented in Tables 1 and 3 make clear how unreliability affects statistical power. The "what if" analysis demonstrates that attenuation of the correlation coefficient due to unreliability compels researchers to have a larger sample size in order to obtain a statistically significant result, in addition to diminishing the potential value of the observed correlation coefficient. This is particularly problematic with small effect sizes (i.e., correlation coefficients).

It should be noted that the "what if reliability" analysis is theoretical in nature because it assumes that the exact correlation coefficient would be replicated in a future study with less (or more) participants. However, this technique is no more theoretical than when one conducts a statistical power analysis in order to determine an appropriate sample size. When conducting power analyses, researchers make the assumption that the hypothesized/desired effect size would be replicated in their future study. Confidence intervals, which have been correctly given elevated status by APA (2001), also represent theoretical values. Consequently, if it is justified to criticize the "what if reliability" analysis for its theoretical interpretation, then other analytic tools (e.g., power analysis, confidence intervals, internal replications) should be held to the same standard. In any case, the proposed "what if reliability" procedure helps to remind researchers the importance of selecting instruments that offer good potential for yielding reliable scores for their intended samples.

Further, it should be noted that the "what if reliability" analysis readily extends to more complex families of the general linear model. For example, squaring the double correction formula above would lead to a disattenuated R^2. Thus, for example, in a multiple regression model, the proportion of variance explained by a regressor variable can be corrected for unreliability, and the role of the sample size investigated using this corrected index.

The "what if" examples provided in this paper show how *p*-values and effect sizes can be potentially misrepresentative of reality when the reliability context is not considered. This, in turn further reinforces the importance of reporting sample-specific reliability estimates for all variables because without these indices, a "what if reliability" coefficient could not be conducted, rendering it impossible for the analyst to rule out unreliability as a rival explanation to the observed findings. Routine reporting of reliability coefficients for scores on all variables within a given study would provide adequate evidence for interested readers to conduct "what-if" analyses if so desired.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

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

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

Guilford, J.P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Henson, R. K. (2000, November). *A primer on coefficient alpha.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

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

Johnson, H. G. (1944). An empirical study of the influence of errors of measurement upon correlation. *American Journal of Psychology, 57*, 521-536.

Johnson, H.G. (1950). Test reliability and correction for attenuation. *Psychometrika, 15*, 115-119.

Kieffer, K. M., & Thompson, B. (1999, November). *Interpreting statistical significance test results: A proposed new "what if" method.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL.

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

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*, 63-75.

Myers, R.H. (1986). *Classical and modern regression with applications.* Boston, MA: Duxbury Press.

Nunnall, 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. (2001, November). *A new proposed binomial test of result direction.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR.

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.

Onwuegbuzie, A.J., & Daniel, L.G. (2001, April). *Indices of score reliability and their applications.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Onwuegbuzie, A.J., & Daniel, L.G. (in press-a). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education.*

Onwuegbuzie, A.J., & Daniel, L.G. (in press-b). 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: Lawrence Erlbaum.

Schmidt, F. L., Hunter, J. E., & Urry V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61*, 473-485.

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

Thompson, B. (1989a). Asking "what if" questions about significance tests. *Measurement and Evaluation in Counseling and Development, 22*(2), 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22*(1), 2-5.

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., 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. (Reprint available through the APA Home Page:

http://www.apa.org/journals/amp/amp548594.html)

Table 1

Corrected values of r based on an observed r = 0.30 given reliability coefficients

Observed reliability coefficient for the Y variable	Observed reliability coefficient for the X variable														
	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1.0
0.3															0.5477
0.35													0.5345	0.5202	0.5070
0.4										0.5477	0.5303	0.5144	0.5000	0.4866	0.4743
0.45									0.5345	0.5163	0.5000	0.4850	0.4714	0.4588	0.4472
0.5							0.5477	0.5262	0.5070	0.4898	0.4743	0.4601	0.4472	0.4352	0.4242
0.55						0.5454	0.5222	0.5017	0.4834	0.4670	0.4522	0.4387	0.4264	0.4150	0.4045
0.6						0.5222	0.5000	0.4803	0.4629	0.4472	0.4330	0.4200	0.4082	0.3973	0.3872
0.65					0.5262	0.5017	0.4803	0.4615	0.4447	0.4296	0.4160	0.4036	0.3922	0.3817	0.3721
0.7				0.5345	0.5070	0.4834	0.4629	0.4447	0.4285	0.4140	0.4008	0.3889	0.3779	0.3678	0.3585
0.75			0.5477	0.5163	0.4898	0.4670	0.4472	0.4296	0.4140	0.4000	0.3872	0.3757	0.3651	0.3554	0.3464
0.8			0.5303	0.5000	0.4743	0.4522	0.4330	0.4160	0.4008	0.3872	0.3750	0.3638	0.3535	0.3441	0.3354
0.85			0.5144	0.4850	0.4601	0.4387	0.4200	0.4036	0.3889	0.3757	0.3638	0.3529	0.3429	0.3338	0.3253
0.9		0.5345	0.5000	0.4714	0.4472	0.4264	0.4082	0.3922	0.3779	0.3651	0.3535	0.3429	0.3333	0.3244	0.3162
0.95		0.5202	0.4866	0.4588	0.4352	0.4150	0.3973	0.3817	0.3678	0.3554	0.3441	0.3338	0.3244	0.3157	0.3077
1.0	0.5477	0.5070	0.4743	0.4472	0.4242	0.4045	0.3872	0.3721	0.3585	0.3464	0.3354	0.3253	0.3162	0.3077	0.3000

Table 2

Minimum sample size needed to obtain a statistically significant "r" based on observed reliability levels for X and Y when observed "r" = .30

Observed reliability coefficient for the Y variable	Observed reliability coefficient for the X variable														
	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1.0
0.3															14
0.35													14	15	16
0.4										14	15	15	16	17	18
0.45									14	15	16	17	18	19	20
0.5							14	15	16	17	18	19	20	21	22
0.55						14	15	16	17	19	20	21	22	23	24
0.6						15	16	18	19	20	21	23	24	25	27
0.65					15	16	18	19	20	22	23	25	26	27	29
0.7				14	16	17	19	20	22	23	24	26	28	29	31
0.75			14	15	17	19	20	22	23	25	27	28	30	31	33
0.8			15	16	18	20	21	23	24	27	28	30	32	33	35
0.85			15	17	19	21	23	25	26	28	30	32	34	35	37
0.9		14	16	18	20	22	24	26	28	30	32	34	36	38	39
0.95		15	17	19	21	23	25	27	29	31	33	35	38	40	42
1.0	14	16	18	20	22	24	27	29	31	33	35	37	39	42	44

Table 3

Excel macro files for the "what if" analysis

	Col. A	Col. B	Col. C	Col. D	Col. E	Col. F	Col. G	Col. H	Col. I	Col. J
Row 1	Computed r =	*	n =	**	Corrected rho =	***	t-calculated =	****	Significance =	*****
Row 2		.05								
Row 3	.05	=IF($B$1>($A$3*B2), " ", ($B$1/(SQRT($A$3)*SQRT(B2))))								
Row 4	.10	=IF($B$1>($A$4*B2), " ", ($B$1/(SQRT($A$4)*SQRT(B2))))								
Row 5	.15	=IF($B$1>($A$5*B2), " ", ($B$1/(SQRT($A$5)*SQRT(B2))))								
Row 6	.20	=IF($B$1>($A$6*B2), " ", ($B$1/(SQRT($A$6)*SQRT(B2))))								
Row 7	.25	=IF($B$1>($A$7*B2), " ", ($B$1/(SQRT($A$7)*SQRT(B2))))								
Row 8	.30	=IF($B$1>($A$8*B2), " ", ($B$1/(SQRT($A$8)*SQRT(B2))))								
Row 9	.35	=IF($B$1>($A$9*B2), " ", ($B$1/(SQRT($A$9)*SQRT(B2))))								
Row 10	.40	=IF($B$1>($A$10*B2), " ", ($B$1/(SQRT($A$10)*SQRT(B2))))								

<u>Legend</u>:

* This is a value entered by the researcher from the original analysis

** This number vacillates (is raised and lowered) until the point at which statistical significance/non-significance is reached

*** This is a value entered from the results of the macros below based on observed reliabilities and original computed r

**** =(F1*(SQRT(D1-2)))/(SQRT(1-(F1*F1)))

***** =TDIST(H1, D1-2, 2)

<u>Note</u>: Once the formulas in Column B have been entered all the way down through "reliability of Y = 1.0", each of the cell formulas may be "dragged" to the right through "reliability of X = 1.0." The designators "$" allow the values for that cell to remain constant, while cell designators which do not have a "$" designator will change to reflect the column for which they are computing values.

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

ERIC®

REPRODUCTION RELEASE

TM033781

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A proposed New "What if Reliability" Analysis for Assessing the Statistical Significance of Bivariate Relationships	
Author(s): Anthony J. Onwuegbuzie, Larry G. Daniel, J. Kyle Roberts	
Corporate Source:	Publication Date: 11/15/01

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	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	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	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	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 end other service egencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Anthony J. Onwuegbuzie	
Organization/Address: School of Education Howard University, 2441 4th St. NW Washington DC 20059	Telephone: 202-249-0416	FAX: 202-249-0417
	E-Mail Address: TonyOnwuegbuzie@aol.com	Date: 11/15/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)