Running head: RELIABILITY GENERALIZATION

Reliability Generalization (RG) Analysis: The Test is Not Reliable

Russell Warne

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association,

New Orleans, February 6, 2008.

Abstract

Literature shows that most researchers are unaware of some of the characteristics of reliability. This paper clarifies some misconceptions by describing the procedures, benefits, and limitations of reliability generalization while using it to illustrate the nature of score reliability. Reliability generalization (RG) is a meta-analytic method developed "to characterize the mean measurement error variance across studies [using a particular instrument] and also the sources of variability of these variances across studies" (Vacha-Haase, 1998, p. 6). Like all meta-analyses, RG uses studies, rather than individuals, as the unit of analysis to help researchers find which study characteristics to predict or explain variablility in observed score reliability coefficients.

Reliability coefficients are among the most important derived from a researcher's data because they form the foundation on which other statistics and any interpretation of scores rest. Unfortunately, there remains a discrepancy between the nature of reliability and the understanding of students and researchers. Most people in the field do not realize that reliability coefficients have their origin in the data they collect and not in tests or instruments. Perhaps this is due to statements in textbooks from authors like Kaplan & Saccuzzo (2005), "The Wechsler adult scale . . . is well constructed and its primary measure—the verbal, performance, and full-scale IQs—are highly reliable. As with all modern tests . . . the reliability of the individual subtests is lower . . ." (p. 270). Statements such as this one only encourage the erroneous belief that reliability comes from tests and not data.

A useful method of correcting this misconception is reliability generalization. To understand and appreciate reliability generalization, however, it is first necessary to understand a few of the traits of reliability coefficients.

Reliability's Importance

In classical test theory, reliability is conceived as the proportion of participants' true score variance to their observed score variance. In every day terms, "reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups" (AERA, APA, & NCME, 1999, p. 25). Due to the complex and intangible nature of the constructs that psychologists measure, all psychological tests yield scores that have some degree of random error. The less error there is in a given administration of a test, the greater the score reliability. Low error—which can be measured through the standard error of measurement or other statistics—means that we can put greater trust and confidence in the scores that a test yields (Thompson, 2006).

Although important in its own right for the information reliability tells us about the scores that we have obtained, score reliability also impacts other aspects of a test's psychometrics. Because score reliability is a necessary but not sufficient condition for score validity, any weakness in score reliability will invariably impact the validity of an instrument's use.

In addition to validity, low score reliability will also dilute the strength of an effect size. As Graham, Liu, & Jeziorski (2006) observed, "Reliability estimates . . . form the upper limit for any effect size using those scores" (p. 704). Henson, Kogan, and Vacha-Haase (2001) added, "Reliability inherently attenuates the maximum possible magnitude of relationship between variables . . . Accordingly, all else being constant, poor score reliability will reduce the power of statistical significance tests" (p. 186). Thus, some of the most important statistics to social scientists—correlational statistics, power (which in turn influences the likelihood of a Type II error), effect sizes, confidence intervals—and the validity of any test scores are influenced directly by the reliability of the data. None of these things can be correctly interpreted without examining the reliability of one's data (Nilsson, Schmidt, & Meek, 2002). Additional consequences of identifying incorrectly the test as having reliability can be found in Thompson (1994).

What Is It That's Reliable?

Given such monumental consequences, it is distressing that misunderstandings about reliability are so common. The most common shortfall in understanding—by far—manifests itself when writers claim that a test is reliable, and not the scores that the test yields (Thompson, 1994). In reality, scores are reliable and not tests (Thompson & Vacha-Haase, 2000) and a test that produces reliable scores for one population may not do so with another (Henson, 2001). Although psychometricians are not unanimous (e.g., Sawilowsky, 2000a, 2000b), the majority of the field seems to agree with Thompson and Vacha-Haase on this one point, (see, for example Capraro & Capraro, 2002; Hanson, Curry, & Bandalos, 2002).

As explained in detail by Thompson and Vacha-Haase (2000), it is incorrect to talk about "the reliability of the test" because (a) there are many types of reliability, (b) reliability inures to the data and not the instrument and, (c) there may be multiple forms of a test (and there always are multiple theoretical parallel forms) or the test may be modified by the researcher (pp. 175-176).

How Tests Aren't Reliable

Thompson and Vacha-Haase's (2000) first point is easy enough to illustrate. One merely has to crack open any of the multitude of textbooks on measurement to see definitions of internal consistency reliability, test-retest reliability, split half and alternate forms reliability, interrater reliability, and others. All of these measure different sources of measurement error and are cumulative (AERA et al., 1999). With so many potential sources of measurement error and ways to measure them, it is obviously foolish to talk about "the reliability" of any test as if reliability were one unitary characteristic.

Thompson and Vacha-Haase's final point states that there may be multiple forms of a test and those may in turn influence validity. This is also easily understandable using common sense. Internal consistency reliability, for example, is highly influenced by test length (Kaplan & Saccuzzo, 2005). Merely choosing the short form of an instrument or administering selected questions is often enough to influence score reliability noticeably (Ross, Blackburn, & Forbes, 2005). Moreover, there may be subtle differences between versions of a test as published by a company (Campbell, Pulos, Hogan, & Murry, 2005). With different forms of a test, the custom modifications that researchers themselves make, and the updates that are produced from time to time, talking about the reliability of *the* test as if any test were one unchanging entity doesn't make much sense.

Reliability Generalization: Showing That Scores Are or Are Not Reliable

By far the most useful tool in showing that reliability is a property of scores and not of tests is RG, a term introduced by Vacha-Haase (1998). Based on validity generalization (Schmidt & Hunter, 1977), the purpose of RG is to

> . . . characterize (a) typical (e.g., mean, median) score reliability for a measure across studies, (b) the variability of score reliabilities, and (c) what measurement protocol features do and do not predict the variability in score reliabilities across administrations. (Henson & Thompson, 2002, p. 114; see also Wallace & Wheeler, 2002, p. 675, for similar thoughts)

By knowing the samples, test forms, or circumstances under which a test yields high reliability, practitioners and researchers can take steps to ensure that the reliability of their data is as high as possible. This, in turn, will improve practitioners' decisions and researchers' confidence in their statistics.

Most people who are skeptical about whether reliability is a property of scores only need to read a few RG studies that show differences in the reliability of the scores of differing samples (e.g., Hellman, Fuqua, & Worley, 2006; Youngstrom & Green, 2003), test forms or formats (e.g., O'Rourke, 2004; Ross et al., 2005), or a combination of these (e.g., Kieffer & Reese, 2002; Vacha-Haase, Kogan, Tani, & Woodall, 2001) to see that talking about a single reliability of a test is overly simplistic and sometimes astoundingly inaccurate.

RG studies also give insight into the subscales of a test, because ". . . consistency of total scores or standard error of measurement may be acceptably high, yet subscores may have unacceptably low reliability" (AERA et al., 1999, p. 31). Several RG studies (e.g., Caruso, 2000; Li & Bagger, 2007; Ryngala, Shields, & Caruso, 2005) have shown that such is the case and that these subscales—and consequentially the entire tests as well—are not measuring their construct or factor with an acceptable level of reliability for a researcher or practitioner's purposes. The scrutiny of a RG is particularly welcomed because subscales are rarely evaluated in the literature by themselves.

Methods of Conducting a Reliability Generalization (RG) Study

As Henson and Thompson (2002) state, "Because RG is not conceived as a monolithic method, there are a variety of ways in which an RG study could be conducted and what variables could be considered in the analysis" (p. 124). With such thoughts in mind, this section explains the most common methods of conducting RG studies and then ends with a few methodological thoughts raised by other authors in order to give the reader a broad understanding of how a typical RG study is conducted. Moreover, the flexibility of RG will become abundantly clear (Henson & Thompson, 2002; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998).

Collecting, Organizing, and Coding Data

After choosing an instrument or set of related tests to evaluate, an RG study begins by casting a wide research net to obtain as much reliability data about a test as possible by entering multiple search terms into all relevant databases. Others augment this by searching for dissertation abstracts that use a test and the reference lists of previously published meta-analyses (e.g., Li & Bagger, 2007), contacting prominent researchers of an instrument to ask for their reliability data (Dunn, Smith, & Montoya, 2006; O'Rourke, 2004), and conducting the study with raw data and a large, diverse sample (Thompson & Cook, 2002).

After eliminating false positives, researchers then examine each relevant article to find whether previous writers reported the reliability coefficients of their own coefficients and whether that information is sufficiently detailed to be included in the meta-analysis. Unfortunately, very few published articles contain reliability information about the data actually used in the study (Vacha-Haase, Kogan, & Thompson, 2000). In fact, frequently fewer than 10% of researchers using an instrument to report their own score reliability coefficients (Barnes, Harp, & Jung, 2002; Beretvas, Meyers, & Leite, 2002; Capraro & Capraro, 2002; Deditius-Island & Caruso, 2002; Vacha-Haase et al., 2001; see also Vacha-Haase, Henson, & Caruso, 2002, p. 564, for a detailed table of reliability coefficient reporting rates). Youngstrom and Green (2003, p. 282), for example, report the embarrassing fact that no researcher who had used the Differential Emotions Scale reported their own reliability coefficients.

Despite the detached writing style typical in quantitative research, the frustration at such low reporting rates is almost palpable in some of these RG articles. Such low reporting rates limits the generalizability of any RG and inhibits the community's ability to ascertain how well a test measures a construct. Moreover, such practices ignore the APA's urging that authors should "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. 596).

However, a few RG researchers have had reporting rates at or above 50% (Leach, Henson, Odom, & Cagle, 2006; Reese, Kieffer, & Briggs, 2002), and there is some indication that reliability coefficients are being reported more in recent years than they have been historically (Dunn et al., 2006; Meier & Davis, 1990). This is good news for those who conduct RG studies in the future.

As depressing as extremely low reliability reporting rates are, an even more egregious act is to cite a previously reported reliability coefficient—often from the test manual—as evidence that one's own data are reliable. Coined "reliability induction" by Vacha-Haase et al. (2000), who define the term as "the practice of explicitly referencing the reliability coefficients from prior reports as the sole warrant for presuming the score integrity of entirely new data" (p. 512).

Deditius-Island and Caruso (2002) expand upon this to describe two types of reliability induction. In the first, researchers cite a test manual or other previously published reliability coefficients and apply the statistics to their own data. In the other (more subtle) type, researchers stay silent on the issue, either not caring about reliability, or quietly assuming that their scores are reliable merely because other researchers' scores have generated acceptable reliability. The former has been christened reliability induction "by report" and the latter reliability induction "by omission" (Shields & Caruso, 2004, p. 256).

Most researchers who engage in this intellectually lazy practice do not justify their behavior by producing evidence that the two samples are comparable (Whittington, 1998). Even when the two samples are similar in composition, previously reported reliability statistics are only useful for comparison purposes to current data. Henson et al. (2001) reminded us that, "Of course, the best evidence of adequate score reliability for one's own data is to actually compute it . . ." (p. 415).

However, all is not lost if a researcher has chosen a test whose literature is lacking reported reliability coefficients. Another option that some (Kieffer & Reese, 2002; Lane, White, & Henson, 2002) have used is to calculate their own reliability estimates through the KR-21 formula (Kuder & Richardson, 1937). This formula is advantageous in generating reliability coefficients for a RG study because the formula merely requires a dichotomous format, the number of participants, the mean of the dependent variable, and either the variance or standard deviation. With this information some unreported reliability coefficients can be estimated. One must realize, though, the KR-21's assumptions of equal item difficulty are rarely perfectly met and KR-21 therefore underestimates reliability (Kaplan & Saccuzzo, 2005). However, the underestimation is usually by .03 or less when reliability is greater than .60 (Lane et al., 2002, p. 688).

After gathering as many reliability coefficients as possible a typical researcher will organize them according to the nature of the test and the data. Most will then examine the ranges and medians of the coefficient of each type of reliability on each scale and subscale.

Coding

The next step that many RG researchers undertake is to code the coefficients they have gathered. Coding systems vary widely, but most are based on the samples that the coefficients refer to and the characteristics of the studies they were published in. Of course, the aspects of the instrument that are of the most interest to the meta-analyst are coded. For example, Ross et al. (2005) examined different reliability coefficients of successive versions of the Patterns of

Adaptive Learning Survey (PALS). Therefore, their coding system took into account which revision of the PALS a sample had been administered.

Analyzing Reliability Coefficients

The first step in analyzing reliability coefficients for some researchers (e.g., (Viswesvaran & Ones, 2000; Yin & Fan, 2000) is to find the average of the reliability coefficients. Feldt and Charter (2006) give seven formulas for doing so and explain quite clearly how to choose the most appropriate one while recognizing that finding any mean of coefficients at all may be unjustified. It is important at this juncture to remind the reader of what was stated above: different reliability coefficients measure different types of measurement error and are cumulative (AERA et al., 1999; Henson & Thompson, 2002). It may not be reasonable to average different coefficients together (Beretvas & Pastor, 2003).

Further analyses through the use of the general linear model are useful for many researchers at this point. Henson and Thompson stated, “Because the general linear model informs us that all classical analyses are fundamentally related, there are many options to examine the relationships between study characteristics (predictor variables) and reliability estimates . . .” (pp. 119-120). As in every step in the RG process, the choice of which method to use is highly flexible. Vacha-Haase (1998), for example, in her pioneering RG used multiple regression and canonical analysis in order to examine the Bem Sex Role Inventory. Henson and Thompson (2002) also mention ANOVA and descriptive discriminant analysis as possible valid methods of analyzing a set of reliability coefficients.

There is some disagreement among those who conduct RG studies as to whether Fisher’s *r*-to-*z* transformation is necessary when performing analyses on reliability coefficients. Henson and Thompson (2002) say that the transformation is unnecessary, while Beretvas et al. (2002) performed the transformation to compensate for the nonnormality of their data. Sawilowsky (2000a), on the other hand, claims that it is always necessary when evaluating any reliability coefficient except internal consistency coefficients. When comparing results from researchers who performed Fisher's transformation to those who didn't, Leach et. al (2006) found no difference between their results. The reader interested in this debate is referred to these articles and other articles about the topic (e.g., Rodriguez & Maeda, 2006).

Reporting Results

The descriptive statistics of the reliability coefficients from prior studies can be reported in a variety of ways, with tables being an especially common method. If a test consists of multiple subscales, some researchers will use boxplots to show the reader the reliability coefficients for each subscale so they can be easily compared. A particularly elegant example of this can be found in Vacha-Haase (2001, p. 50) where several reliability coefficients of the MMPI's clinical subscales are presented in an easily understandable way that allows the reader to quickly compare the different distributions.

In accordance with the APA's recognition that "confidence intervals combine information on location and precision and . . . are, in general, the best reporting strategy" (APA, 2001, p. 22), Henson and Thompson (2002) recommended using confidence intervals in RG studies. However, their recommendation has not caught on among other researchers, despite the publication of Excel formulas to calculate RG confidence intervals easily (Vacha-Haase et al., 2002).

Problems with Reliability Generalization Studies

The most obvious problem with RG lies with the limited amount of data that goes into the meta-analysis. With low rates of reporting reliability coefficients being the rule rather than the exception, it is hard to ascertain how representative the reported coefficients really are. As mentioned previously, low reliability reporting stunts the generalizability of an RG study. Concerning this problem, Thompson and Vacha-Haase (2000) noted, “We may not like the ingredients that go into making this sausage, but the RG chef can only work with the ingredients provided by the literature” (p. 184).

This presents a cromulent example of the “file drawer problem” (Rosenthal, 1979) where results from published studies may not accurately represent studies done as a whole. Many RG researchers openly acknowledge this problem (e.g., Caruso, 2000). Additionally, low reporting even in published studies would only exasperate this distortion. The intrepid researchers who have made an effort to obtain unpublished reliability coefficients have produced mixed results as to whether published research presents a skewed picture of the true nature of the reliability estimates that test data produce. Li and Bagger (2007) found that published studies do have higher reliability coefficients than unpublished studies, but O’Rourke (2004) found no difference between the unpublished reliability coefficients he obtained from researchers after contacting them personally and the coefficients that had been published.

Another limitation of RG studies lies in the fact that Cronbach’s (1951) alpha is by far the most reported measure of reliability (Hogan, Benjamin, & Brezinski, 2000), probably because alpha only requires one administration of a test, whereas almost all other measures require multiple administrations. For example, Nilsson et al. (2002) only found one test-retest reliability coefficient in their exhaustive RG literature research on the Career Decision-Making Self-

Efficacy Scale. Consequentially, their report (and the vast majority of RG studies) only gives the reader an assessment of the internal consistency of an instrument.

Miscellaneous Benefits of Reliability Generalizations

In addition to showing that score reliability is a property of scores and not of tests, RG can teach psychometricians much about reliability and its relationship to tests. For example, although conventional folk wisdom in the field declares that longer tests are more reliable, such is not always the case (Kieffer & Reese, 2002). Another study found that homogenous groups who were administered the Coopersmith Self-Esteem Inventory yielded higher reliability coefficients than heterogeneous groups (Lane et al., 2002), contrary to the tendency for higher reliability to come from more heterogeneous samples. Such nuggets of research that run counter to intuition are valuable because they force us to re-examine our assumptions about reliability.

Conclusion

Despite problems with generalization and other issues (Dimitrov, 2002), RG is clearly a highly useful method that illustrates with empirical data several characteristics of reliability. First, tests are not reliable; scores are. Such is shown in the varying reliability coefficients that are obtained with different samples, test forms, test lengths, test versions, and even test administration conditions (language, etc).

Second, RG helps researchers and practitioners better understand the instruments that they use in their work. By knowing which populations a test is best for or which subscale(s) have unacceptably low reliability practitioners can make informed decisions when examining test scores, and researchers and test developers will know which instrument is most appropriate under a given set of circumstances.

Third, RG studies help researchers obtain scores that will produce the largest effect sizes and most power. Moreover, RG encourages researchers to report their own reliability in the future and to not commit the “sin” of reliability induction (Thompson & Vacha-Haase, 2000, p. 179).

In short, ten years after Vacha-Haase opened the floodgates of reliability generalization, the benefits are clear. Though not pinned down to one set of set of procedures, the researchers who conduct RG meta-analyses are giving the psychological community a better understanding of the instruments they use every day.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Reliability and errors of measurement. In *Standards for educational and psychological testing* (pp. 25-36). Washington, DC: American Educational Research Association.

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

Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*, 603-618.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*, 570-589.

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*, 75-95.

Campbell, J. S., Pulos, S., Hogan, M., & Murry, F. (2005). Reliability generalization of the Psychopathy Checklist applied in youthful samples. *Educational and Psychological Measurement, 65*, 639-656.

Capraro, R. M., & Capraro, M. M. (2002). Myers-Briggs Type Indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement, 62*, 590-602.

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

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

Deditius-Island, H. K., & Caruso, J. C. (2002). An examination of the reliability of scores from Zuckerman's Sensation Seeking Scales, Form V. *Educational and Psychological Measurement, 62*, 728-734.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*, 783-801.

Dunn, T. W., Smith, T. B., & Montoya, J. A. (2006). Multicultural competency instrumentation: A review and analysis of reliability generalization. *Journal of Counseling & Development, 84*, 471-482.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66*, 215-227.

Graham, J. M., Liu, Y. J., & Jeziorski, J. L. (2006). The Dyadic Adjustment Scale: A reliability generalization meta-analysis. *Journal of Marriage and Family, 68*, 701-717.

Hanson, W. E., Curry, K. T., & Bandalos, D. L. (2002). Reliability generalization of Working Alliance Inventory Scale scores. *Educational and Psychological Measurement, 62*, 659-673.

Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A reliability generalization study on the Survey of Perceived Organizational Support: The effects of mean age and number of items on score reliability. *Educational and Psychological Measurement, 66*, 631-642.

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

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

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

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

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological Testing*. Belmont, CA: Thomson Wadsworth.

Kieffer, K. M., & Reese, R. J. (2002). A reliability generalization study of the geriatric depression scale. *Educational and Psychological Measurement, 62*, 969-994.

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

Lane, G. G., White, A. E., & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement, 62*, 685-711.

Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66*, 285-304.

Li, A., & Bagger, J. (2007). The Balanced Inventory of Desirable Responding (BIDR): A reliability generalization study. *Educational and Psychological Measurement, 67*, 525-544.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113-115.

Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement, 62*, 647-658.

O'Rourke, N. (2004). Reliability generalization of responses by care providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement, 64*, 973-990.

Reese, R. J., Kieffer, K. M., & Briggs, B. K. (2002). A reliability generalization study of select measures of adult attachment style. *Educational and Psychological Measurement, 62*, 619-646.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Ross, M. E., Blackburn, M., & Forbes, S. (2005). Reliability generalization of the Patterns of Adaptive Learning Survey goal orientation scales. *Educational and Psychological Measurement, 65*, 451-464.

Ryngala, D. J., Shields, A. L., & Caruso, J. C. (2005). Reliability generalization of the Revised Children's Manifest Anxiety Scale. *Educational and Psychological Measurement, 65*, 259-271.

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

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

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

Shields, A. L., & Caruso, J. C. (2004). A reliability induction and reliability generalization study of the Cage Questionnaire. *Educational and Psychological Measurement, 64*, 254-270.

Thompson, B. (1994, January). *It is incorrect to say, "The test is reliable": Bad language habits can contribute to incorrect or meaningless research conclusions.* San Antonio, TX: Southwest Educational Research Association. (ERIC Document Reproduction Service No. 367707)

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: The Guilford Press.

Thompson, B., & Cook, C. (2002). Stability of the reliability of LibQUAL+ scores: A reliability generalization meta-analysis study. *Educational and Psychological Measurement, 62*, 735-743.

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., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562-569.

Vacha-Haase, T., Kogan, L. R., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement, 61*, 45-59.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522.

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

Wallace, K. A., & Wheeler, A. J. (2002). Reliability generalization of the life satisfaction index. *Educational and Psychological Measurement, 62*, 674-684.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21-37.

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

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

Youngstrom, E. A., & Green, K. W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement, 63*, 279-295.