DOCUMENT RESUME

ED 445 077 TM 031 754

| | |
|---|---|
| AUTHOR | Cousin, Sherri L.; Henson, Robin K. |
| TITLE | What Is Reliability Generalization, and Why Is It Important? |
| PUB DATE | 2000-01-27 |
| NOTE | 28p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000). |
| PUB TYPE | Reports - Descriptive (141) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | Data Analysis; *Error of Measurement; *Estimation (Mathematics); *Generalization; Meta Analysis; *Reliability |

ABSTRACT

Researchers consistently fail to report reliability estimates for data used in their studies. This lack of reporting hinders appropriate evaluation and interpretation of data and may lead to inappropriate conclusions. Because reliability is inured to scores obtained from a test, and not the test itself, it is important to report score reliability in both measurement and substantive studies. Failure to report reliability estimates is a practice that plagues social science research and may undermine results across studies. Reliability generalization (RG) is a meta-analytic method that looks at the reliability of scores from tests and helps to determine what is causing measurement error across studies. RG has the potential to help researchers use specific tests more accurately and to help identify specific sample characteristics and other factors that influence and affect score data. The purpose of this paper is to discuss the sample dependency of score reliability, emphasize the importance of measurement, even in substantive studies, and explain the premises and procedures of RG. (Contains 1 figure, 1 table, and 29 references.) (Author/SLD)

Running head: RELIABILITY GENERALIZATION

# What is Reliability Generalization, and Why is it Important?

Sherri L. Cousin and Robin K. Henson

The University of Southern Mississippi

Paper presented at the annual meeting of the Southwest Educational Research Association, January 27 ,2000, Dallas. Correspondence concerning this manuscript can be sent to the second author at, Department of Educational Leadership and Research, Box 5027, Hattiesburg, MS 39401-5027; or via email at robin.henson@usm.edu.

## Abstract

Researchers consistently fail to report reliability estimates for data used in their studies. This lack of reporting hinders appropriate evaluation and interpretation of data and may lead to inappropriate conclusions. Because reliability is inured to scores obtained from a test, and not the test itself, it is important to report score reliability in both measurement and substantive studies. Failure to report reliability estimates is an ill practice that plagues socio-science research and may undermine results across studies. Reliability Generalization (RG) is a meta-analytic method that looks at the reliability of scores from tests and helps to determine what is causing measurement error across studies. In short, RG has the potential to help researchers more accurately utilize specific tests and help to identify specific sample characteristics and other factors that influence and affect score data. The purpose of this present paper is to: (a) discuss the sample dependency of score reliability, (b) emphasize the importance of measurement, even in substantive studies, and (c) explain the premises and procedures of Reliability Generalization.

## Tests are Not Reliable

One of the prevailing yet misleading beliefs among researchers is that tests are reliable. In fact, it is scores, and not tests, that are either reliable or unreliable (Thompson, 1995; Thompson & Daniel, 1996). The false presumption that tests are reliable has led to an underreporting of reliability estimates and a general dismissal of its relevance in research altogether. Reliability Generalization is an important meta-analytic method that may serve to correct this ill-practice, for its application highlights the relevancy and usefulness of reliability coefficients and links its importance to data interpretation. The purpose of the present paper is to: (a) discuss the sample dependency of score reliability, (b) emphasize the importance of measurement, even in substantive studies, and (c) explain the premises and procedures of Reliability Generalization, a new meta-analytic method that examines sources of measurement error variance across studies.

Invariably, even a cursory examination of the literature reveals researchers attesting that their "test is reliable" or making mention of the "reliability of [their] test" (Vacha-Haase, 1998). Thompson (1994) stated that this "language is both incorrect and deleterious in its effects on scholarly inquiry, particularly given the pernancious consequences that unconscious paradigmatic beliefs can exact" (p. 839). The inappropriate use of language, and subsequent confusion about the issue, is encouraged when researchers claim that their "test is reliable" without calculating the exact reliability on their data. Thompson (1992) warned that,

> this is not just an issue of sloppy speaking—the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. (p. 436)

For example, the recent report from the APA Task Force on Statistical Inference (Wilkinson & the Task Force on Statistical Inference, 1999) presents contradictory information concerning score reliability. Correctly, the report stated, "It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees" (p. 597). Unfortunately, the very next paragraph incorrectly begins with "[b]esides showing that an instrument is reliable . . ." (p. 597).

This inconsistency highlights the pervasive nature of misconceptions surrounding score reliability. Even well-respected textbooks present misleading information. For example, Kerlinger (1992) defined reliability as ". . . the accuracy or precision of a measuring instrument"(p. 405). Further, to illustrate, he gives the example of a sportsman, who, in attempting to determine the firing accuracy of two guns, places the guns on two fixed bases. Then, they are "zeroed in" by a sharpshooter and equal rounds are fired at matching targets for comparison (p. 405). However Thompson (1994), Vaccha-Hasse (1998) and others would rebuff this example, for measuring instruments can yield both reliable and unreliable scores and this definition, like others, fails to address what other factors (e.g., wind velocity, accuracy of the sharpshooter, etc.) could have impacted the results as well. Hence, although tests are important and do impact scores, they do not do so exclusively. Other factors influence score results and must be considered by researchers as well.

This "unconscious paradigmatic belief" to which Thompson (1994, p. 839) alluded assumes that tests, in and of themselves, are responsible for score reliability. The assumption follows that if a test yielded reliable scores in the past, then certainly it will yield reliable scores in the future. The assumption is faulty. It has repeatedly been argued that reliability is a function of scores and not tests (cf. Pedhazer & Schmelkin, 1991; Thompson, 1994; Vacha-Haase, 1998). Accordingly, the scores obtained from a given test are dependent, at least to some degree, on the characteristics of the sample tested. As sample characteristics change, it is very possible (even likely) that score reliability estimates will fluctuate even when using the same test under identical testing conditions. (It is important to note here that other issues, such as testing conditions, may also impact score reliability.) Reinhardt (1996) explained that "[b]oth the characteristics of the person sample selected and the characteristics of the test item can affect coefficient alpha" (p. 6).

To illustrate that reliability (coefficient alpha) is heavily impacted by sample characteristics, a brief hypothetical example is employed here. Assume that a test was given to two samples, one of which was very homogeneous on the trait tested, and the other being very heterogeneous on the trait of interest. It is intuitive that these two samples would possess very different total score distributions. The homogeneous sample, no doubt, would have a smaller total score variance than the heterogeneous group. After all, if persons in the heterogeneous group actually differed greatly on the trait of interest, then their total scores would vary considerably (assuming the test was well designed).

Table 1 presents scores for six persons on five items for these two types of samples. Items are scored right (1) or wrong (0). As theoretically expected, these data yielded a larger total score variance and coefficient alpha for the heterogeneous group ($\underline{s}^2 = 3.50$; $\alpha = .83$) than the homogeneous group ($\underline{s}^2 = .30$; $\alpha = -5.00$).

INSERT TABLE 1 ABOUT HERE

As noted by Thompson (1994), "The same measure, when administered to more heterogeneous or more homogeneous sets of subjects, will yield scores with differing score reliability" (p. 839). Reliability, then, is not inured to tests, but to other influences, including sample characteristics. The data in Table 1 illustrate this possibility. Cronbach's coefficient alpha (1951) differed in the example because alpha is impacted by the ratio between the sum of the item variances and the total test variance. Generally, as item variance decreases and total variance increases, reliability estimates will increase. As such, tests given to heterogeneous groups are likely to yield more reliable scores than tests given to people who are common on the trait of interest. Readers are referred to Reinhardt (1996) and Thompson (1999) for detailed treatments of this ratio including explanations of why alpha can have a negative value. In brief, however, the negative value is a product of the alpha formula itself (below) and is paradoxical since alpha is a variance-accounted-for statistic in a squared metric.

$$\alpha = \left(\frac{n}{n-1}\right) \frac{SD_t^2 - \Sigma(SD_i^2)}{SD_i^2}$$

Theoretically, scores are considered more reliable when they correctly order persons on the trait of interest. If persons with more of the trait (e.g., knowledge of math) score higher than persons with less of the trait, then the scores are considered accurate (reliable). In test-retest situations, the theory is clearly illustrated. If Susie, for example, really is knowledgeable at math, then she should score high in both testing conditions. If she does not, then her scores are considered less reliable (accurate). As such, the total score variance plays an important role in coefficient alpha estimates. The greater the total score variance, the more likely the sample is ordered correctly on the trait of interest, thereby honoring the theoretical assumptions of score reliability.

Since reliability may vary (and likely will) from sample to sample, then it is important that researchers report reliability estimates for their data in all studies. Unfortunately, many researchers fail to cite or even recognize the importance of reporting score reliability (Thompson, 1994; Vacha-Haase, 1998). This problem is quite pronounced. In a review of the common reporting practices of three major research journals over a period of 7 years (1990-1997), Vacha-Haase (1999) reported that 64% of the articles did not provide reliability coefficients for their data, and one-third of the articles ignored the issue altogether by not even mentioning reliability of scores. When reliability is not reported, the reader is left to guess whether the scores used in the current study were reliable. Since scores, and potentially reliability, will change with future samples, references to score reliabilities in test manuals or prior studies do not help the matter. Out of the many articles that do not report reliability for the data in hand, one may begin to question the stability of reliability across studies for a given test. For example, Vacha-Haase (1998) found considerable variability in reliability coefficients for the Bem Sex Role Inventory (Bem, 1974) and noted that "absent an expectation that authors will routinely report reliability coefficients for their data, one is left to wonder if a 'file drawer' problem regarding such results might mean that reliability coefficients . . . [are] even more variable [than observed]" (pp. 14-16). Willson (1980) stated 20 years ago that the lack of score reliability reporting was "inexcusable at this late date" (p. 9).

## The Importance of Measurement

Researchers who do not report the reliability of their scores are making assumptions about their data and, as a result, are potentially drawing unwarranted and misguided conclusions that would otherwise change if they examined score reliability in the context of their data. The problem is perhaps attributable to the perceived irrelevancy of measurement issues in sociobehavioral research. Some time ago, Kuder (1941) observed that measurement is nearly disregarded. As researchers, measurement issues are not uniformly taught nor is the importance of score reliability demonstrated prior to interpreting and reporting results in the literature. As Pedhazur and Schmelkin (1991) noted,

> Measurement is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. . . . It is, therefore, not surprisingly that little or no attention is given to properties of measures used in many research studies. (pp. 2-3)

Consequently, there exists a systemic problem in socioscience research in which the failure to interpret findings in light of score reliability coefficients may adversely impact conclusions.

Nunnally (1982) cautioned against this practice when he noted that "if we rely solely on the testing instrument, or how the researcher used the test, the reality of results is limited" (p. 1589).

Importantly, score reliability "is critical in detecting effects in substantive research" (Reinhardt, 1996, p. 3). If, for example, two variables are correlated and either of the variables was measured in a manner that yielded perfectly unreliable scores, then the resulting effect size (i.e., squared correlation) is inescapably zero. At this point, of course, the correlation cannot be statistically significant, regardless of the size of the sample and observed statistical power. Outside of perfectly unreliable scores, the degree that measurement problems will affect effect sizes can be estimated. Locke, Spirduso, and Silverman (1987) noted that "the correlation between scores from two tests cannot exceed the square root of the product for reliability in each test" (p. 28). As Reinhardt (1996) observed,

> Thus, if a researcher is correlating scores having a reliability of .9 with scores having a reliability of .6, the correlation cannot exceed .73. Prospectively, researchers must select measures that will allow the detection of effects at the level desired; retrospectively, researchers must take reliability into account when interpreting findings. (p. 3)

When one set of scores has a .50 alpha and the second set of scores has .60 estimate (which is probably not terribly uncommon), the attenuation is worse, limiting the possible correlation to .55 [$\underline{r}^2$ = .30]. In this case, only a little over one-half of the variance in the dependent variable is going to be predictable, regardless of the true relationship between the variables. Certainly, measurement properties must be considered in result interpretation.

Yet, it is common practice not to examine results in light of reliability. When trying to replicate findings, researchers may arduously set up a matching study with similar conditions using the same dependent variables, independent variables, and sample size, and ponder over why their results did not replicate. However, if the researchers had reported score reliabilities, they could have observed not only the conditions under which the results may be replicable, but also the potential attenuation of replicability due to poor score reliability that impacted their results.

Vacha-Haase (1999) emphasized that "score reliability is critically important, even in substantive studies . . . because the score reliability of the data being analyzed directly affects . . . results and possible interpretations of such results" (p. 335). It is very possible that some studies draw certain conclusions when, if score reliability were considered, the results would have painted a different picture altogether. As noted by Pedhazur and Schmelkin (1991), assuming reliability is inured to the test is "inappropriate and potentially misleading" (82). The problem is exacerbated by ignoring the fact that scores, not tests, account for reliability or unreliability. Keep in mind that "a single instrument can produce scores which are reliable or unreliable....[Thus], reliability is a property of scores on a test for a particular group of examinees" (Rowley, 1976, p. 53). Researchers, before drawing conclusions, must (not should) report and consider score reliability. As Thompson (1990) emphasized, "[M]easurement integrity is critical to the derivation of sound research conclusions" (p.585). In other words, if we ignore what is influencing the reliability of our scores in research, then the inferences that we are drawing from our findings are potentially flawed.

## Reliability Reporting and Journal Policies

Fortunately, there has been some, albeit little, movement in the field regarding the need to report reliabilities. While the dialogue concerning the issue is not new, actual reporting practices have changed little. Furthermore, it is unlikely that reporting practices will change unless the editorial policies of journals mandate that reliability be reported for the data in hand. As gatekeepers of what research gets recognized, journal editors should examine their policies regarding score reliability and emphasize (even require) that reliability estimates be reported and that results be interpreted in light of such estimates. At present, several journal, including Educational and Psychological Measurement, Journal of Applied Psychology, and Journal of Experimental Education, either require or strongly encourage as part of their acceptance criteria, that score reliability be reported along with effect sizes (Vacha-Haase, 1999). This growing trend will hopefully encourage further emphasis on measurement properties and promote customary use of this reporting practice in the field.

## Vacha-Haase's Reliability Generalization

In addition to journal policies emphasizing the relevancy of score reliability and its link to data interpretation, interest in score reliability has increased due to Reliability Generalization (RG). Vacha-Haase (1998) introduced RG as a new meta-analytic method that explores sources of measurement error across studies which employ the same instrument. RG examines the "mean measurement error variance across studies and . . . the sources of variability of these variables across studies" (Vacha-Haase, 1998, p. 7). RG is akin to the renown validity generalization method (see Hunter & Schmidt, 1990; Schmidt & Hunter, 1977). In fact, "the same premises and methods [used in validity generalization] can be applied to study score reliability" (Vacha-Haase, 1998, p. 9).

In RG, the study becomes the unit of analysis, and the reliability estimate becomes the dependent variable. It is then possible to explore variation of reliability estimates for a given instrument across studies and to determine what study characteristics (e.g., sample size, test form, etc.) can predict the variation. Unlike validity generalization, which examines to what "degree . . . validity obtained in one situation can be generalized in another without further study" (Rafilson, 1991, p.

1), RG is a meta-analytic technique that assesses the measurement error variance and, importantly, defines for the researcher where these variances may lie within the sample (Vacha-Haase, 1998). In other words, conducting an RG study provides information about what is potentially causing measurement error across studies using an instrument. For instance, length of the test could predict reliability variation. Typically, as a test increases in length, it yields more reliable scores. However, as Thompson (1990) noted, when discussing the Bem Sex Role Inventory (BSRI, Bem, 1974), the short version (20 items) generally yields more reliable scores on the feminine scale than the long version (40 items). Vacha-Haase (1998) empirically validated this observation through an RG study of the BSRI. RG helps to determine these types of critical factors that impact scores across studies and enables researchers to ultimately determine what may affect score reliability when a particular instrument is used. Further, with these types of influences identified, RG may help researchers draw more accurate conclusions and interpretations about score results.

## What is Reliability Generalization?

In reliability generalization, the first step is to assemble studies that utilize a specific instrument in analyzing and computing results, such as the BSRI that Vacha-Haase (1998) used in her introductory article describing the RG method. Of course, any achievement or attitudinal instrument could be utilized. As a meta-analytic method, RG does necessitate that enough studies using the instrument of choice have been published to justify synthesis of research. For example, Vacha-Haase (1998) located 628 articles utilizing the BSRI.

Once the studies are collected, the articles are analyzed for correctly reported reliability estimates. It is important to separate the articles in categories of reporting because researchers vary in how and whether they report reliability coefficients. In the test case, Vacha-Haase (1998) separated her studies into three categories. The first category of studies (65.76%) reported no reliability; the second category reported reliability estimates from the test manual or prior studies (14.65%) or mentioned reliability as being reported elsewhere but specific coefficients were not reported (6.53%); the third category of studies did calculate reliability estimates for the researched data and reported coefficients (13.06%). The Reliabilities from the third category of studies were used as the dependent variable in the RG analysis (Vacha-Haase, 1998). It is important to emphasize that of the 628 studies using the BSRI, 546 (roughly 87%) <u>did not report</u> reliability coefficients for the data in hand. Since they failed to provide reliability for the data in hand, these studies were excluded from the RG analysis.

Once studies are identified, the reported articles are then coded for specific and <u>well thought-out</u> "study characteristics" that potentially can affect the variance of scores across the studies. For instance, it would be expected that if you gave a psychological profile test to a group of depressed patients (homogenous), and you administered the same test to a group of college students (heterogeneous), the results of the depressed patients would yield less variant scores, and thus lower score reliability, than the heterogeneous group. Therefore, when determining "study characteristics", it is vital to determine what could be impacting and creating errors across studies. Other examples of characteristics include, but are not limited to, population types (e.g., clinical v. general) type of scale used; (e.g., Likert), language used, length of test, type of reliability coefficient used, and a variety of sample characteristics such as ethnicity, age, education, and gender.

## Coding Data for RG

Once determined, the selected characteristics are dummy coded. An important consideration is to code for characteristics that capture contrast in the data. Remember, RG explores errors across studies and potential causes for those errors. Hence, an example of coding for RG could be by gender: male (0), female (1), and mixed (2). In addition, population type could be coded as a clinical population (1) and a non-institutionalized population (0). Other examples include: non-degreed participants (1) v. degreed participants (0); blacks (1) v. non-Blacks (0); 7-point Likert scale (1) v. 5-point Likert scale (0). Again, coding should be based on the type of instrument used and what differences one is attempting to make known. Vacha-Haase's (1998) study examined 11 different study characteristics.

Once all study characteristics are properly identified and coded, regression (or some other general linear model analysis) is then run to determine which of the identified study characteristics are related to variation in reported reliability estimates, and thereby identify potential sources of measurement error for test across studies. Importantly, both beta weights and structure coefficients are then examined to determine the best predictors of reliability variation, thereby assessing what study characteristics tend to contribute to measurement error. When an instrument consists of more than one subscale, then it may be preferable to use a multivariate analysis such as canonical correlation. Reliability estimate fluctuation can (and should) also be examined descriptively, perhaps through a box and whisker plot or some other graphical representation. In Vacha-Haase's (1998) BSRI study, the data told two stories in the box and whisker plot: (a) the reliability coefficients for the Feminine scores were higher than those for the Masculine scores, and (b) for both scales, reliability coefficients were "fairly variable across studies" (p. 14). For example, Figure 1 graphically presents results of a hypothetical RG study for a self-referent test containing three subscales: self-esteem, self-image, and social-self. The RG study examined 15 different articles using the three subscales.

INSERT FIGURE 1 ABOUT HERE

In this example, scores on the self-esteem variable generally had high reliabilities. Estimates for the self-esteem subscale were also relatively homogeneous across studies. For self-image, there was a considerable range of reliabilities reported across studies (.49 to .89). Also, variability of estimates was greater for scores from the self-image subscale than the self-esteem variable. One would have less confidence that the self-image subscale would always yield reliable scores. Finally, the range of score reliabilities for the social self variable was similar to the self-image subscale (.50 to .90). However, the social-self subscale exhibited the greatest variability between the 25th and 75th percentile. Again, as with self-image, one should use caution and not assume that the social-self subscale would always yield reliable scores. Of course, one should never assume such. For instance, even for the self-esteem variable, one of the studies reported a marginal alpha of .50, despite the fact that the subscale generally yielded acceptable score alphas in other studies (see outlier observation in the boxplot).

## Why is Reliability Generalization Important?

In Vacha-Haase's (1998) initial RG study, several study characteristics predicted variation in the reliability estimates, including the type of coefficient used (alpha or KR v. test-retest), sample size, and test length. The long form of the test tended to produce less reliable scores than the short form of the test for the Feminine scale. Of course, some study characteristics were not predictive (e.g., 5 v. 7 point Likert scale). These results are important not only for those considering the use of the BSRI instrument in the future, but also regarding the use of RG as a method to examine score reliability across studies. RG may very well foster a greater understanding of various testing instruments and what influences results when employing them. Furthermore, RG adds to the growing support that reliability coefficients must and should be considered, analyzed, and reported. As a meta-analytic method, RG spotlights the considerable variation in score reliability that can occur across studies using the same instrument. Understanding what influences on score reliability may also guide researchers in creating better tests that yield more reliable scores. For instance, the confirmation that the short form of the Feminine scale of the BSRI tends to yield more reliable scores is significant. And, equally important, the use of a 5 v. 7 point Likert scale had no impact on score reliability on the BSRI. But, do such results necessarily hold true for other instruments? Of course not. Accordingly, as researchers, it is crucial that we investigate these and other potential factors which could impact score reliability. RG provides a means of doing just that.

While Vacha-Haase's (1998) original RG method only included studies that reported reliabilities for the data in hand, a recent extension of the method also includes those studies that fail to report reliability but do report the mean and standard deviation for the scale of interest. Using the KR-21 reliability formula, Henson, Kogan, and Vacha-Haase (2000) estimated score reliability for those studies employing the Teacher Efficacy

Scale (Gibson & Dembo, 1984) that did not report reliability for the data in hand (see Kuder & Richardson, 1937 for discussion of the KR-21 formula.) If KR-21 is used to estimate reliability, then RG can then include a great many more studies in its analysis. Theoretically and as a caution, KR-21 "may be expected to give an underestimate of the reliability coefficient in situations not favorable for its application" (Kuder & Richardson, 1937, p. 159). The newest extension of RG will be able to empirically evaluate this assumption as well as more closely evaluate the impact of sample characteristics (e.g., homogeneity v. heterogeneity) on score reliability.

Interested readers are also referred to an upcoming issue of Educational and Psychological Measurement (EPM, Vol. 60, slated for April 2000) in which several RG studies will be published. Given the infancy of RG as a method, the apparent utilization of the approach is encouraging. The upcoming EPM issue includes three RG studies, one on the Beck Depression Inventory and two concerning the NEO personality scales (cf. Caruso, 2000, Viswesvaran, 2000; Yin & Fan, 2000). In addition, the issue will present a critique of RG as a method (Sawilowsky, 2000) and a thoughtful response by EPM's editor and Vacha-Haase (Thompson & Vacha-Haase, 2000).

In short, RG informs researchers about reliability variation across studies using a particular instrument. It identifies what factors are related to measurement error within either the test itself or the sample used. Moreover, RG has the potential to allow for better construction and administration of tests and may facilitate more thorough and complete discussions of score reliability in the literature.

## References

Bem, S. L. (1974). The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology, 42, 155-162.

Caruso, J. C. (2000). Reliability Generalization of the NEO personality scales. Educational and Psychological Measurement, 60.

Gibson, S., & Dembo, M. (1984). Teacher efficacy: A construct validation. Journal of Educational Psychology, 76, 569-582.

Henson, R. K., Kogan, L., & Vacha-Haase, T. (2000, April). A reliability generalization study of the Teacher Efficacy Scale and related instruments. Paper to be presented at the annual meeting of the American Educational Research Association, New Orleans.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Kerlinger, F. (1992). Foundations of behavioral research. Fort Worth: Harcourt.

Kuder, G. F. (1941). Presenting a new journal. Educational and Psychological Measurement, 1, 3-4.

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

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (1987). Proposals that work: A guide for planning dissertations and grant proposals (2nd ed.). Newbury Park, CA: Sage.

Nunally, J. C. (1982). Reliability of measurement. Encyclopedia of Educational Research, 1589-1601.

Pedhqzur, E. j., & Schmelkin, L. P. (1991). Measurement, design and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rafilson, F. (1991). The case for validity generalization. ERIC, 1-4.

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

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Sawilowsky, S. S. (2000). Psychometrics vs datametrics: Comment on Vacha Haase's "Reliability Generalization" method and some EPM editorial policies. Educational and Psychological Measurement, 60.

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

Thompson, B. (1990). ALPHAMAX: A program that maximizes coefficient alpha by selective item deletion. Educational and Psychological Measurement, 50, 585-589.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

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

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525-534.

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

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. Educational and Psychological Measurement, 56, 197-208.

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

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. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. Journal of Experimental Education, 67, 335-341.

Viswesvaran, V. C. (2000). Measurement error in "big Five Factors" of personality assessment: Reliability Generalization across studies and measures. Educational and Psychological Measurement, 60.

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

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9, 5-10.

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

Table 1

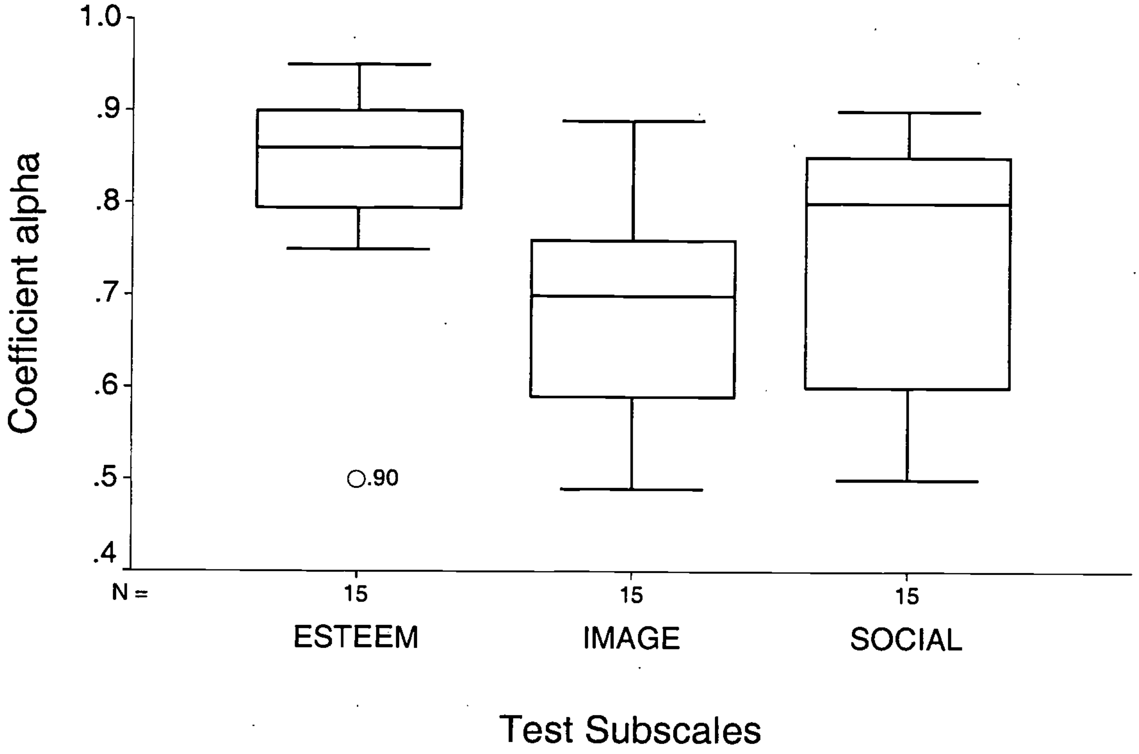
Hypothetical Data for Heterogeneous and Homogeneous Samples

| Person/ Statistic | Item 1 | 2 | 3 | 4 | 5 | Total Score |
|---|---|---|---|---|---|---|
| | | | Heterogeneous Sample | | | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 1 | 1 | 0 | 0 | 0 | 2 |
| 4 | 1 | 1 | 1 | 0 | 0 | 3 |
| 5 | 1 | 1 | 1 | 1 | 0 | 4 |
| 6 | 1 | 1 | 1 | 1 | 1 | 5 |
| Item $s^2$ | .14 | .22 | .25 | .22 | .14 | |
| Total $s^2$ | | | | | | 3.50 |
| alpha | | | | | | .83 |
| | | | Homogeneous Sample | | | |
| 1 | 0 | 1 | 0 | 1 | 0 | 4 |
| 2 | 1 | 0 | 1 | 0 | 1 | 3 |
| 3 | 0 | 1 | 0 | 1 | 0 | 4 |
| 4 | 1 | 0 | 1 | 0 | 1 | 3 |
| 5 | 0 | 1 | 0 | 1 | 0 | 4 |
| 6 | 1 | 0 | 1 | 0 | 1 | 3 |
| Item $s^2$ | .25 | .25 | .25 | .25 | .25 | |
| Total $s^2$ | | | | | | .30 |
| alpha | | | | | | -5.00 |

Note. This illustration is adapted from Reinhardt (1996) and Thompson (1999).

# Figure 1:

## Hypothetical RG Study Results



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

ERIC®

TM031754

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: What is Rellability Generalization, and Why is it Important?

Author(s): Sherri L. Cousin and Robin K. Henson

Corporate Source: The University of Southern Mississippi

Publication Date: January 27, 00

## 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 foplevel 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
e,→
ase

| | |
|---|---|
| Signature: Sherri L. Cousin | Printed Name/Position/Title: Sherri L. Cousins |
| Organization/Address: The Univ. of Southern Mississippi - Educational Leadership<br>Hattiesburg, MS 39406 Chairman, Ric Keaster | Telephone: 601-296-0110 / FAX: 601-296-0110<br>E-Mail Address: / Date: January 27, 00 |

## 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:

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**
**1100 West Street, 2nd Floor**
**Laurel, Maryland 20707-3598**

**Telephone: 301-497-4080**
**Toll Free: 800-799-3742**
**FAX: 301-953-0263**
**e-mail: ericfac@inet.ed.gov**
**WWW: http://ericfac.piccard.csc.com**