DOCUMENT RESUME

ED 455 294 TM 033 148

AUTHOR	Henson, Robin K.; Thompson, Bruce
TITLE	Characterizing Measurement Error in Test Scores across Studies: A Tutorial on Conducting "Reliability Generalization" Analyses.
PUB DATE	2001-04-14
NOTE	33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Guides - Non-Classroom (055) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Coding; *Error of Measurement; Psychometrics; *Reliability; *Scores; Test Results; Tutoring
IDENTIFIERS	*Generality

ABSTRACT

Given the potential value of reliability generalization (RG) studies in the development of cumulative psychometric knowledge, the purpose of this paper is to provide a tutorial on how to conduct such studies and to serve as a guide for researchers wishing to use this methodology. After some brief comments on classical test theory, the paper provides a practical framework for structuring an RG study, including: (1) test selection with an eye toward frequency of test use and reporting practices by authors; (2) development of a coding sheet that will capture potential variation in score reliability across studies; (3) procedural recommendations regarding data collection; (4) identification and use of potential dependent variables; and (5) application of general linear model analyses to the data. (Contains 40 references.) (SLD)

Running head: RG TUTORIAL

Characterizing Measurement Error in Test Scores Across Studies: A Tutorial on Conducting "Reliability Generalization" Analyses

Robin K. Henson

University of North Texas 76203-1337

Bruce Thompson

Texas A&M University 77843-4225

and

Baylor College of Medicine

Paper presented at the annual meeting of the American Educational Research Association, April 14, 2001, Seattle. The first author may be contacted at rhenson@tac.coe.unt.edu. The second author may be contacted via his web page, http://acs.tamu.edu/~bbt6147/.

Abstract

As emphasized by the recent APA Task Force on Statistical Inference report (Wilkinson & APA Task Force on Statistical Inference, 1999), reliability can vary for scores across different administrations of a given instrument. Consistent with this view, Vacha-Haase (1998) proposed reliability generalization methodology to examine the variance of measurement error across studies, and therefore provide meta-analytic information regarding the ability of a test to yield reliable scores upon repeated administrations. Given the potential value of reliability generalization studies in the development of cumulative psychometric knowledge, the purpose of the present paper is to provide a tutorial on how to conduct these studies, and therefore serve as a guide for researchers wishing to employ this methodology.

Characterizing Measurement Error in Test Scores Across Studies: A Tutorial on Conducting "Reliability Generalization" Analyses

As noted by Gronlund and Linn (1990), "Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. Thus it is more appropriate to speak of the reliability of 'test scores' or the 'measurement' than of the 'test' or the 'instrument'" (p. 78, emphasis in original). This view is echoed in the recent report of the APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999), which recommended that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (p. 596).

Part of the logic for reporting score reliability in all studies relates directly to the Task Force's mandate to also report effect sizes, because "Interpreting the size of observed effects requires an assessment of the reliability of the scores" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596). As Reinhardt (1996) observed,

> Reliability is critical in detecting effects in substantive research. For example, if a dependent variable is measured such that the scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results

> will not be statistically significant at any sample size, including an incredibly large one. (p. 3)

Accordingly, appropriate interpretation of observed effects should invoke an examination of the reliability of the obtained scores.

Unfortunately, as Henson, Kogan, and Vacha-Haase (in press) noted, "the incorrect but common phraseology concerning the 'reliability of the test' leads many to incorrectly assume that reliability inures to tests rather than scores. . . ." This misperception fails to honor the fact that an instrument may yield scores of varied reliability upon repeated administrations (cf. Crocker & Algina, 1986; Dawis, 1987; Gronlund & Linn, 1990; Pedhazur & Schmelkin, 1991; Thompson 1994; Vacha-Haase, 1998). Thompson and Vacha-Haase (2000) discussed the "etiology of [this] endemic misspeaking about reliability" and noted the dangers of using "the phrase 'the reliability of the test' as a telegraphic shorthand in place of truthful but longer statements (e.g., 'the reliability of the test scores')" (p. 178).

Rather than present the reliability of obtained scores, researchers at times report coefficients from test manuals or previous studies as relevant for their data. However, as Pedhazur and Schmelkin (1991) noted, "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 study under consideration" (p. 86). Vacha-Haase, Kogan, and Thompson (2000) called the use of prior coefficients for present data "reliability induction," because this decision reflects a generalization of a specific instance (e.g., test manual coefficient) to a more general state of affairs (e.g., future scores from the test).

This process may be legitimate when the reliability induction occurs for a sample that is comparable to the inducted sample in terms of composition and variability. As Crocker and Algina (1986) observed,

> Reliability is property of the scores on a test for a particular group of examinees. Thus, potential test users need to determine whether reliability estimates reported in test manuals [or prior studies] are based on samples similar in *composition* and *variability* [italics added] to the group for whom the test will be used. (p. 144)

Dawis (1987) emphasized the need for future samples to match normative samples as regards reliability: "Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the *intended target population* - an obvious but sometimes overlooked point" (p. 486, italics added). Unfortunately, few researchers make explicit comparisons of sample compositions and variability when inducting reliability estimates (Vacha-Haase et al., 2000).

Furthermore, it is not uncommon for authors to even fail to report such basic information as total score variability. For example, Henson et al. (in press) observed that variability was reported in only 27.7% of studies also reporting reliability for the data in hand on four measures of teacher self-efficacy and locus of control. Similarly, Capraro, Capraro, and Henson (in press) observed a 51.4% reporting rate for the Mathematics Anxiety Rating Scale. Whittington (1998) presented an empirical evaluation of how well educational researchers report their measures and noted some unfortunate deficits.

Reliability Generalization

Because reliability can (and will) vary across test administrations, Vacha-Haase (1998) extended the concept of validity generalization (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977) and proposed a meta-analytic methodology called reliability generalization (RG) to characterize measurement error in a test's scores across studies. RG can be used to (a) describe the variability of reliability estimates across studies, (b) identify study characteristics which may be predictive of reliability variance, and (c) provide cumulative psychometric knowledge regarding study factors that influence score reliability estimates. Among other possibilities, researchers using RG results may be able to tailor their studies to maximize reliability; experimentally manipulate study characteristics and examine impact on reliability; and examine the practical relationships between reliability, effect sizes, power, and statistical significance in published research.

One of the first RG studies appeared in MECD, in which Helms (1999) reported a meta-analysis of Cronbach alphas from the White Racial Identity Scale. Other RG studies have been conducted, and include examinations of the Bem Sex Role Inventory (Vacha-Haase, 1998), Beck Depression Inventory (Yin & Fan, 2000), "Big Five Factors" of personality across various tests (Viswesvaran & Ones, 2000), NEO-Five Factor Inventory (Caruso, 2000), Teacher Efficacy Scale and related instruments (Henson et al., in press), and Mathematics Anxiety Rating Scale (Capraro et al., in press). Table 1 presents a listing of these and other RG studies to date.

INSERT TABLE 1 ABOUT HERE

Purpose

Of course, the potential benefits of RG studies will only be realized if such investigations continue to be conducted and allotted space in the published literature. In the interest of facilitating this outcome, the purpose of the present paper is to provide an accessible tutorial on how to conduct RG studies. After some brief comments on classical test theory, the present paper provides a practical framework for structuring an RG study, including (a) test selection with an eye toward frequency of test use and reporting practices by authors, (b) development of a coding sheet that will capture potential variation in score reliability across studies, (c) procedural recommendations regarding data collection, (d) identification and use of potential dependent variables, and (e) application of general linear model analyses to the data.

The discussion is not intended to be exhaustive, and there are no doubt reasonable applications or extensions of RG methodology that are not addressed here. Although Vacha-Haase (1998) introduced one approach, RG is not a monolithic method, and may use "different analytic tools" (p. 16). As Thompson and Vacha-Haase (2000) noted, ". . .we do not see RG as involving always a single genre of analyses" (p. 185). Accordingly, conducting an RG study is seemingly limited by two things: (a) the creativity and insightfulness of the researcher and (b) the information reported in the studies examined.

A Brief Note on Classical Test Theory

Comprehensive discussions of classical test theory (CCT) are given by Crocker and Algina (1986) and classic works such as Gulliksen (1950) and Lord and Novick (1968). Nevertheless, several major points regarding CCT will be emphasized here.

According to CCT, a person's observed score (X_O) can be conceptualized as the additive function of his or her true score (X_T) and error score (X_E):

$$X_O = X_T + X_E.$$

Furthermore, a person's true score represents "the average score a person would obtain on an infinite number of parallel forms of a test, assuming that the person is not affected by taking the tests (i.e., assuming no practice or fatigue effect)" (Hopkins, 1998, p. 114). Of course, the true score is primarily theoretical in nature.

When we extend CCT to a sample of scores, we are typically concerned with the degree that the observed scores represent true score variance. It is logical that if all the variance in the scores was due to random responding, then no one person's score could be considered accurate, or reliability measured, and the total variance of the scores would be solely due to error fluctuations. Accordingly, CCT reliability estimates follow the general formula,

$$\sigma_{TRUE}^2 + \sigma_{ERROR}^2 = \sigma_{TOTAL}^2,$$

such that the percentage of reliable variance in a set of scores is given as the ratio of true score variance to total score variance:

$$r_{tt} = \sigma_{TRUE}^2 / \sigma_{TOTAL}^2.$$

From this general formula for CTT score reliability, it is clear that the reliability estimate is modeled as a ratio of variances, and therefore, is in a squared, variance-accounted-for metric. This ratio is a measurement analog of the more familiar eta^2 and $\underline{R}^2$ uncorrected effect sizes derived in general linear model analyses (Dawson, 1999; Snyder & Lawson, 1993).

Central to estimates of score reliability is the notion of total score variance. Total score variation is assumed to be an indicator of the degree of accuracy in the scores. As Henson et al. (in press) explained:

> In terms of classical measurement theory (holding the number of items on the test and the sum of item variances constant [for internal consistency estimates]), increased variability of total scores suggests that we can more reliably order people on the trait of interest, and thus more accurately measure them. This assumption is made explicit in the test-retest reliability case, when consistent ordering of people across time on the trait of interest is critical in obtaining high reliability estimates.

For test-retest reliability, if the order of participants (or the ranking of who possesses most to least of the trait) changes across testing occasions, then our reliability of measuring these participants is less than ideal. Therefore, reliability estimates hinge largely on the variance of the total test scores. As this variance increases, the reliability estimate will also tend to increase (holding all else constant), due to greater theoretical confidence that we have accurately ordered (measured) the participants on the trait. When total variance is small, "the closeness of the scores will mean that small random fluctuations will alter rank orderings" (Thompson & Vacha-Haase, 2000, p. 178). This dynamic is directly related the present discussion because, as will be noted, one facet of RG studies seeks to capture study characteristics that speak to the degree of total score variance (i.e., group homogeneity versus heterogeneity) in the sample.

Selecting a Test for an RG Study

Essentially, any test yielding scores for which reliability can be calculated can be submitted to an RG study. However, the test must enjoy enough use in the published literature to allow for a meta-analytic synthesis. In reality, what matters is not the frequency of test use, but rather the frequency of reliability reporting when the test was used. This is because it is the reliability coefficient that typically becomes the dependent variable in RG studies (see below for other dependent variable options).

Because reliability is a function of the observed scores, then the relevant reliability estimates are those for the data in hand, rather than coefficients reported from test manuals or prior studies. Unfortunately, few researchers actually report reliability for their obtained scores. Yin and Fan (2000) observed a 7.5% reporting rate for the Beck Depression Inventory. Caruso (2000) reported a 15.2% rate for the NEO personality scales. Others have observed higher levels of appropriate reporting but rates are unlikely to exceed very far beyond 40%. A little more than twenty years ago, Willson (1980) noted that "Only 37% of the [American Educational Research Journal] studies explicitly reported reliability coefficients for the data analyzed" (p. 8). Little has changed since Willson's review.

If authors fail to report these coefficients, then the RG researcher left without the central piece of RG information and it really doesn't matter if there are 500 published articles using the test. As Thompson and Vacha-Haase (2000) metaphorically explained,

> . . . it is important to remember that RG studies are a meta-analytic characterization of what is hoped is a population of previous reports. 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. Obviously, at some extreme the literature may not be sufficient to conduct an RG study, just as can

> happen in both "validity generalization" and in substantive meta-analysis studies. (p. 184)

Accordingly, some preliminary screening may be warranted when identifying a test. This could be handled in variety of ways but one option would be to at least semi-randomly sample articles using the test and examine them for the type of information reported. The RG researcher may then be able to generalize expectations about additional articles before expending too much effort.

Finally, because RG requires pouring over multiple articles seeking fairly specific information, it certainly helps to orient oneself toward a test that has some intrinsic interest to the researcher. Much like writing a dissertation, academic motivation can be tested after fully immersing oneself to the project, particularly when the number of articles to review is high!

Developing a Coding Form

Once a test has been identified, a coding form can be developed which will be used to record study characteristics from the articles employing the test. It is important that the nature of the test be considered when developing the coding form. Because classical reliability estimates are largely a function of total score variance, issues of group homogeneity become central to potential fluctuations in reliability estimates. Therefore, the astute RG researcher will seek to identify study characteristics that will potentially capture differences in group homogeneity.

For example, in Yin and Fan's (2000) study of the Beck Depression Inventory, the authors chose to code contrasts to determine whether the sample represented (a) clinical psychiatric patients versus non-clinical respondents, (b) respondents with physical diseases versus those without physical disease, and (c) respondents with a substance addiction versus those with no addiction. Other study features were coded, but these few illustrate how the study characteristics examined should stem for the test purpose and the nature of the samples used.

Of course, the number and type of study characteristics that can be coded is largely limitless in theory. In practice, as noted above, the RG researcher is in fact limited by the information presented by the literature. A well-crafted variable is of little use if information about the variable is never reported.

Some examples of study characteristics that have been used in prior RG studies include: (a) sample size; (b) type of Likert scale used; (c) gender, coded a variety of ways including use of a measure of gender homogeneity; (d) number of items on the test; (e) test form; (f) type of reliability coefficient reported; (g) clinical versus non-clinical samples; (h) teaching level; (i) years teaching experience; (j) age; (k) sample ethnicity; (l) English versus non-English test version; (m) self-referent versus other-referent, (n) standard deviation; and (o) many others. Again, the relevancy of these or other characteristics stems from the test itself and its typical usage. The RG researcher should reflect on these issues and thoughtfully develop variables of interest.

A few practical matters are worth noting. First, limit the coding sheet to one page if possible, thereby eliminating the harassment of manipulating multiple sheets. Second, if the variable is continuous in nature, then the information can be directly used. If the variable is categorical in nature (e.g., gender, ethnicity), then some form of coding (e.g., dummy coding) will be required prior to analysis. Third, seek to collect as much (relevant) information from the articles as possible. A practical reality is that although the researcher may believe a variable will be frequently reported, the researcher may in fact be dead wrong and find little useable information. In such cases, other variables must be used. It is wise to have sufficient variables recorded, lest the researcher finds him or herself with no useable data after many hours of examining articles. Typically, the researcher would not be terribly eager to re-examine the articles to collect more data!

Procedural Recommendations

Finding Relevant Articles

A primary procedural issue in an RG study is the identification of articles employing the test in question. Common article databases (e.g., PsycINFO, PSYCLIT, or ERIC) provide a useful means of developing the pool of articles. Importantly, an RG study would normally seek to identify all of the articles using the test and therefore represent the current population of article use. Accordingly, the search terms used in the database should be sufficiently broad to capture the various derivations of the test nomenclature, including abbreviations. For example, in an RG study of the NEO personality scales, Caruso (2000, p. 240) searched the PsycINFO database and retained citations containing the terms "NEO," "five-factor model," "FFI," or "five factor inventory."

As a caution, however, overly broad searches may result in too many false hits, or articles identified by the database but which did not use the test. The number of false hits can be large (cf. Capraro et al., in press: 261 false hits; Yin & Fan, 2000: 762 false hits) which results in time spent weeding out the imposters. (Interpreted with a positive spin, finding a large number of false hits would also mean there would eventually be fewer articles to actually read and code, which takes considerably more time than simply noting that an article does not use the test!)

Once the citations have been identified, the articles must be obtained, read, and coded. Any given library or electronic journal service is unlikely to house every citation needed. Therefore, efforts should be made to attain these articles via services such as inter-library loan. Because the intent is to capture all articles using the test, these procedures should be thorough and well documented. Researchers' varied stylistic preferences will guide the actual data collection procedures followed. Some will make copies of all articles and examine them at remote locations; others will prefer to read the articles in libraries or via electronic resources.

Categorizing Articles

Typically, the articles using the test of interest can be grouped into at least three categories: (a) articles that make no mention of reliability, (b) articles that report reliability only from previous studies or test manuals, and (c) articles that report reliability for the data in hand. Of course, other categories are possible, such as articles that "simply stated that the reliability was acceptable" (Caruso, 2000, p. 240).

Because reliability is a function of scores and not the test per se, only the articles reporting reliability for the data in hand are relevant for a typical RG study. This inherent limitation makes article screening important, because the RG researcher does not wish to examine all of his or her articles only to find that there is simply not enough information to warrant a meta-analysis. Of course, the researcher could point out the gross failure to report reliability for the test.

Organize, Organize, Organize

The need to stay organized in this process cannot be overemphasized. The number of articles is often large and decisions regarding whether the articles are useful need to be clearly recorded. Of course, there are many ways to handle this type of data collection. One possibility that has been successfully utilized by the first author (and others) is to simply create a master list of all the citations found in the database searches. This master list is then printed and cut so that each citation is on its own slip of paper. These citations can then be physically categorized. Citations for which articles were copied or a code sheet was completed can be attached to the article/code sheet to avoid confusion. Finally, all of the categorized and coded citations can be double-checked against the master list to verify that all articles have been found (or that reasonable attempts were made to find the article) and record decisions regarding each article's utility in the RG study.

Organization becomes even more important when more than one person is involved with data collection (e.g., graduate students). Developing a well-thought out plan prior to data collection can tremendously reduce effort and confusion later.

RG Dependent Variables

Classical Reliability Coefficients

As introduced by Vacha-Haase (1998), RG studies have typically employed the reliability estimate itself as the dependent variable in analyses. Continuous or coded study/sample variables are used in some fashion to predict the variance in the reliability estimates. These estimates may be internal consistency or test-retest coefficients, and both have been used in previous studies. It should be noted, however, that in the CTT framework, these estimates do not measure the same measurement error. Instead, the error due to item sampling (internal consistency) and the error due to time (test-retest) are separate and cumulative; they do not represent the same source of measurement error simply measured two different ways. (Anastasi & Urbina, 1997; Henson, 2000b; Thompson, 1991b).

Internal consistency estimates are found more regularly in the literature as they only require one administration of the test. Furthermore, as demonstrated in prior RG studies (Capraro et al., in press; Viswesvaran & Ones, 2000; Yin & Fan, 2000), test-retest coefficients tend to be lower than internal consistency estimates (a finding that also suggests that these two type of coefficients model different sources of measurement error). Therefore, the joint use of these types of coefficients in a single analysis may be questionable, because the effects of the independent variables (study characteristics) would be confounded with the type of reliability coefficient. When sufficient coefficients exist, separate analyses could be conducted to evaluate whether study characteristics differentially predict the two types of estimates. Yin and Fan (2000) observed this dynamic in their RG study.

To Transform or Not to Transform

Sawilowsky (2000) suggested that test-retest coefficients should be submitted to Fisher's _r_-to-_z_ transformation before they can legitimately be used in analyses because they are "ordinary Pearson product-moment coefficients of correlation" and therefore "suffer from the skewed distribution problem" (p. 169). While it is generally understood that measures of internal consistency (e.g., KR-20, coefficient alpha) are in an $\underline{r}^2$ type metric (and therefore are in a continuous scale), the unsquared Pearson _r_ is used as the test-retest coefficient of stability.

Thompson and Vacha-Haase (2000) explained, however, that even though we use the unsquared Pearson _r_ to model stability reliability, what we are estimating is still the population variance-accounted-for statistic. And, we do so by invoking the concept of and correlation between parallel forms (or the same test administered twice). As Lord and Novick (1968) explained,

> The square of the correlation between observed scores and true scores is equal to the correlation between parallel measurements. Thus, assuming at least one pair of parallel measurements can be obtained, we have succeeded in expressing and unobservable quantity $\rho_{XT}{}^2$ in terms of $\rho_{XX'}$, a parameter of a (bivariate) observed-score distribution. (pp. 58-59)

Thompson and Vacha-Haase (2000) noted, then, that "often the way we estimate score reliability is by computing unsquared *r* values. But by doing so, nevertheless what we are estimating is variance-accounted-for universe values (i.e., reliability coefficients)" (p. 186, emphasis in original). Therefore, we believe the Fisher *r*-to-*z* transformation is not necessary before submitting reliability coefficients to analysis.

Standard Error of Measurement

Measurement-related statistics other than reliability coefficients could reasonably be used in RG studies. One immediately obvious possibility would be to use the standard error of measurement (SEM) as the dependent variable. We model the SEM as the product between the observed standard deviation and the square root of the measurement error, $\underline{SD}_X\ (1 - \underline{r}_{XX})^{.5}$. Conceptually, this is an estimate of the degree we would expect a person's score to fluctuate due to random fluctuations of error.

Importantly, however, the SEM estimates an individual's observed score variation in the population, rather than the reliability of the set of scores. Furthermore, the SEM assumes that each person contributed equally to measurement error across the distribution. In fact, this is not an accurate assumption as persons who score above the mean tend to have positive errors of measurement and persons who score below the mean tend to have negative errors of measurement (Hopkins, 1998). If the test is targeted toward the average test-taker, then persons who score in the tails of the distribution will tend to have more measurement error reflected in their scores. Accordingly, use of the SEM to create confidence intervals for scores across the distribution is "rather crude" (Thompson & Vacha-Haase, 2000, p. 187). The confidence bands actually should be wider for scores in the distribution tails.

These limitations notwithstanding, the SEM could be calculated by RG researchers, assuming reliability and standard deviation are reported, and treated as their dependent variable (cf. Yin & Fan, 2000). As noted, the expectation, however reasonable, that both of these pieces of information will be reported in the same study may often not be met. Furthermore, use of the SEM requires that the same form of the test be used in all applications to make comparisons the SEM across studies reasonable.

General Linear Model Analyses

There are a variety of methods that could be invoked to analyze RG data. Because the general linear model informs us that all classical analyses are fundamentally related (cf. Bagozzi, Fornell & Larcker, 1981; Cohen, 1968; Henson, 2000a; Knapp, 1978; Thompson, 1991a), there are many options to similarly examine the relationships between study characteristics (predictor variables) and reliability estimates or other measurement statistics (criterion variables). All of these analyses (e.g., ANOVA, multiple regression, discriminant analysis, canonical correlation analysis), yield $\underline{r}^2$ type effect sizes (e.g., eta^2, $\underline{R}^2$) that should be both reported and interpreted (Snyder & Lawson, 1993). Furthermore, the contributions of individual variables to the synthetic, linear combination of the variables (e.g., Yhat scores in regression, function scores in canonical analysis) should be evaluated by examining <u>both</u> the weights (e.g., betas) and structure coefficients (Courville & Thompson, 2001; Thompson & Borrello, 1985). Examination of the standardized weights only may lead to inappropriate interpretations of predictor importance.

Vacha-Haase (1998) demonstrated this flexibility and used both multiple regression and canonical correlation analysis in her initial RG study. Canonical analysis was employed because reliability estimates for scores on both the masculine and feminine subscales of the Bem Sex Role Inventory were simultaneously examined as dependent variables. The point here is that RG analyses can reasonably differ, and that the analysis should match the purpose and needs of the research questions.

Summary

RG has emerged as a potentially important advancement in the measurement literature. RG can be used to characterize the measurement error of scores for a given test across repeated administrations and identify study characteristics that may be predictive of measurement error variation. 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 analyses.

As a final thought, RG may also be useful for exploring measurement error of scores across constructs, rather than individual tests. Viswesvaran and Ones (2000) demonstrated this possibility when examining the "Big Five Factors" of personality assessment. Somewhat differently, Henson et al. (in press) examined teacher self-efficacy and locus of control across four measures. These studies allow for integration not only of reliability data for scores from a given test, but also reliability information across tests presuming to measure similar constructs. While ultimately different, these types of studies resemble the concepts of convergent and discriminant validity discussed by Campbell and Fiske (1959).

As noted, the ultimate benefit of RG studies will only be realized if researchers continue to engage in this type of research and if the published literature recognizes their work. As regards the former of these two issues, the purpose of the present paper has been to provide a brief tutorial on conducting these analyses, and offer important references the reader may consult for more information.

References

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

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Capraro, M. M., Capraro, R. M., & Henson, R. K. (in press). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. Educational and Psychological Measurement, 61.

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

Caruso, J.C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J. (in press). Reliability of scores from the Eysenck Personality Questionnaire: A Reliability Generalization (RG) study. Educational and Psychological Measurement, 61.

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

Courville, T. & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. Educational and Psychological Measurement, 61, 229-248.

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

Dawis, R. V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

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

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

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

Helms, J.E. (1999). Another meta-analysis of the White Racial Identity Attitude Scale's Cronbach alphas: Implications for validity. Measurement and Evaluation in Counseling and Development, 32, 122-137.

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

Henson, R. K. (2000b, November). A primer on coefficient alpha. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC Document Reproduction Service No. forthcoming)

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

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

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

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

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

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

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

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

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.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

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

Thompson, B. (1991b). A review of generalizability theory: A primer by R. J. Shavelson & N. W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

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

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

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

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

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

Vacha-Haase, T., Tani, C.R., Kogan, L.R., Woodall, R.A., & Thompson, B. (in press). Reliability Generalization: Exploring reliability variations on MMPI validity scale scores. Assessment.

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.

Wilkinson, L., & American Psychological Association (APA) 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)

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

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.

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

Table 1

Illustrative RG Studies

Author(s)	Year
Capraro, Capraro, and Henson	(in press)
Caruso	(2000)
Caruso, Witkiewitz, Belcourt-Dittloff, and Gottlieb	(in press)
Helms	(1999)
Henson, Kogan, and Vacha-Haase	(in press)
Vacha-Haase	(1998)
Vacha-Haase, Kogan, Tani, and Woodall	(2001)
Vacha-Haase, Tani, Kogan, Woodall, and Thompson	(in press)
Viswesvaran and Ones	(2000)
Yin and Fan	(2000)

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

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Characterizing Measurement Error in Test Scores Across Studies: A Tutorial on Conducting "Reliability Generalization" Analyses

Author(s): Henson, Robin K. and Thompson, Bruce

Corporate Source: University of North Texas

Publication Date: April 2001

II. REPRODUCTION RELEASE:

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

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

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

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

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

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

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

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

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

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

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

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

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

Sign here,→ please

Signature:

Printed Name/Position/Title: Robin K. Henson / Assist. Professor

Organization/Address: Dept. of Tech. & Cog. P.O. Box 311337
Denton, TX 76203-1337

Telephone: 940-369-8385

FAX: 940-565-2185

E-Mail Address: rhenson@tac.coe.unt.edu

Date: 6/11/01

(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:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
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
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