The Validity of Reliability Generalization

Michael C. Rodriguez
University of Minnesota

April 2000

A paper presented in the symposium "Meta-Analysis: Issues and Practice"
at the annual meeting of the American Educational Research Association, New Orleans, LA.

Recently, *Educational and Psychological Measurement* published a special issue on "reliability generalization," (April 2000). The issue contained both a critique and a defense of the procedure as well as three applications studies. The authors of the critique and defense, as well as the applications, relied on classical test theory for their conceptualization of reliability and error variance. They also, unfortunately, abbreviated their methods regarding the actual synthesis of reliability coefficients to the extent that the methodology remained ambiguous. Not only was the statistical framework applied in the reliability generalization studies ambiguous, they appeared to be weak in their treatment of several standard issues facing the research synthesist. Although references were made to the methods used in validity generalization as a parallel framework, several important issues were ignored.

This paper is offered to uncover some of the underlying strengths of research synthesis (meta-analysis) as an effort to provide a stronger framework for continued work in the synthesis of psychometric coefficients. This paper is also a result of work completed through participation in SynRG (Synthesis Research Group at Michigan State University) under the direction of and with thoughtful comments from Betsy Becker. I begin with several aspects of reliability theory that were downplayed in Thompson and Vacha-Haase's defense of reliability generalization and Sawilowsky's (2000) critique. I briefly comment on each article in the recent special edition of

Educational and Psychological Measurement (Vol. 60, No. 2). I then briefly describe a few critical methodological issues in meta-analysis that were apparently ignored in the recent applications of reliability generalization. Finally, I present an IRT perspective to the concept of reliability generalization offered by Mark Reckase (personal communication, March 2000). In part, the following discussion critically questions the validity of reliability generalization. Moreover, it presents a case for a more structured and principled approach to continued efforts in the synthesis of psychometric coefficients.

Reliability, Revisited

In the classical test theory framework, the observed score is the sum of two independent components: the true score and error score. The error score is a random variable that may have associated with it any number of factors, including the physical measurement process itself, characteristics of the environment under which measurements were obtained, and temporal changes in any number of individual-based characteristics. Lord and Novick (1968) suggested that the degree to which these factors are controlled and randomized determines different true scores and thus different error (residual) scores. "For each definition of true score, of course, we have a different error score. Just what is included in the error score depends entirely on the conditions under which measurements are made" (p. 39).

Reliability has been conceived of differently in several frameworks based on the mathematical model used in the scoring and analysis of scores from educational and psychological instruments. Traditionally, "the *reliability of a test* is defined as the squared correlation ρ_{XT}^2 between observed score and true score. From the relation $\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$ we see

that *the reliability of a test is a measure of the degree of true-score variation relative to observed-score variation*" (Lord & Novick, 1968, p. 61).

This relationship is also commonly written as $\rho_{XX} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$. It is conceivable that two estimates of reliability are equal where each is based on different levels of error variance with proportional changes in observed score variance (Reckase, personal communication, March 2000). Similarly, different estimates of reliability could result from changes in true score variance with no concomitant changes in error score variance. In fact, under the IRT framework describe below, changes in "reliability" clearly result from changes in the ability or trait (theta) distribution.

Focusing on the theoretical conceptualization of the reliability estimate as a correlation, Samejima (1977) argued that the resulting estimate of reliability not only depends on the test itself, but also on the specific group of examinees, as the case with any correlation.

> To give an extreme example, however refined the test may be, the reliability coefficient is zero if all examinees have exactly the same true score. Conversely, it is easy to make a poorly constructed test look good by calculating the correlation coefficient for a group of examinees whose ability levels are substantially different from one another. (p. 233)

Because of this, generalizability is limited. Similarly, the functional relationship between test length and reliability was illustrated in the Spearman Brown prophecy formula; however, not entirely prophetic. The relationship between test length and reliability has been well documented. Lord and Novick (1968) demonstrated that the reliability of a test of infinite length is unity and of zero length is zero.

"The situation is substantially different in latent trait theory where the standard error of estimation does have an intrinsic meaning, since test information function is defined independently of any specific group of examinees" (Samejima, 1977, p. 234).

The limitations of the classical test theory conception of reliability have been debated for nearly 40 years. These limitations, in part, were the driving force behind the efforts to develop strong test theory, including IRT. Lord and Novick (1968) cautioned researchers early on regarding overuse or abuse of resulting reliability estimates. "The reader should note that without further information, the reliability coefficient along is of little value for describing a test as a measuring instrument. The reason is that a large reliability coefficient can often be obtained by administering the test to a significantly heterogeneous group of examinees" (p. 199).

The recent edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) reiterated this comment: "The reporting of reliability coefficients alone, with little detail regarding the methods used to estimate the coefficient, the nature of the group from which the data were derived, and the conditions under which the data were obtained constitutes inadequate documentation" (p. 31). The *Standards* described the critical information necessary to interpret reliability information appropriately, including the identification of major sources of error, descriptive statistics regarding the size of errors, and where possible, generalizability information based on relevant dimensions of the measurement procedures such as forms, scorers, and administrations. In addition, the examinee population from which the above descriptions were derived must be described, "as the data may accurately reflect what is true of one population but misrepresent what is true of another" (p. 27).

Measurement error is infrequently defined in a substantive way by researchers and, as argued by Schmidt and Hunter (1999), " the processes that produce measurement error are not

mysterious" (p. 192). Three substantive processes leading to measurement error that are important in psychological measurement include random response error, transient error, and specific factor error. Different estimates of reliability assess different sources of error. Schmidt and Hunter are particularly interested in estimating the magnitude of error for the purpose of control—the failure of researchers to control for the biases introduced in their research because of measurement error has slowed the accumulation of knowledge. In their work in validity generalization and elsewhere, they have advocated the correction for biases due to measurement error, range restriction, and other sources (Hunter and Schmidt, handbook). "It is not possible to have accurate empirical tests of theories and hypotheses unless the biases introduced into data by measurement error are controlled and correct for" (p. 183).

While discussing issues regarding the nature of specific factor errors of measurement, Schmidt and Hunter (1999) presented the following example. Suppose the literature contains several verbal ability measures that are not classically parallel, constructed at different times by different researchers. We choose to define verbal ability as the common factor among these measures and administer five scales with the total combined score as the final observed score. The appropriate reliability of these scores could be estimated with coefficient alpha as though each scale were an item in a five-item instrument. The result could be considered the generalizability coefficient as conceived of by Cronbach, Gleser, Nanda, & Rajaratnam (1972). They suggested that the results would "generalize to the population of all such non-parallel measures of verbal ability" (p. 196) and result in better theory construction with greater generality.

Reliability Generalization

Although the synthesis of reliability coefficients is not a new activity (see, for example,), Vacha-Haase has brought the concept of reliability generalization to the forefront. The issue has raised so much interest that *Educational and Psychological Measurement* has devoted space in a special issue to several papers on the topic. In addition, an AERA mini-course was offered at the 2000 annual meeting on the methodology involved in reliability generalization (after the completion of this paper).

Meta-Analysis, Revisited

In their introduction to the *Handbook of Research Synthesis*, Cooper and Hedges (1994) defined research syntheses as attempts to "integrate empirical research for the purpose of creating generalizations" (p. 5). This implies evaluating the limits and "modifiers" of resulting generalizations. Three additional contributions of research syntheses include a critical analysis of the research involved, attempts to resolve conflicts in the literature, and identification of potential future research agendas.

Two common questions addressed by the research synthesist are: "How confident can we be that the findings can be generalized beyond a small subset of populations, settings, and procedures?" and "Does the research advance the theoretical understanding of a phenomenon?" (Hall, et al., 1994, p. 18). Through a research synthesis, we can empirically evaluate the validity of generalizations by testing moderator variables believed to be associated with certain populations of subjects, settings, or administration procedures involved in primary studies.

One issue often overlooked by the research synthesist regards the universe to which generalizations are made. This addresses the issue related to fixed effects versus random effects inferences and statistical analyses. Hedges (1994) adequately described the differences:

> In the fixed effects, or conditional, model, the universe to which generalizations are made consists of ensembles of studies identical to those in the study sample except for the particular people (or the primary sampling units) that appear in the studies. Thus, the studies in the universe differ from those in the study sample only as a result of the sampling of people into the groups of the studies. The only source of sampling error or uncertainty is therefore the variation resulting from the sampling of people into studies. (p. 30).

> In the random effects, or unconditional, model, the study sample is presumed to be literally a sample from a hypothetical collection (or population) of studies. The universe to which generalizations are made consists of a population of studies from which the study sample is drawn. Studies in this universe differ from those in the study sample along two dimensions. First, the studies differ from one another in study characteristics and in effect size parameter. …Second, in addition to differences in study characteristics and effect size parameters, the studies in the study sample also differ from those in the universe as a consequence of sampling of people into the groups of the study. (p. 31).

The argument to support the use of random effects analysis suggests that the studies we observe in the literature are, more or less, accidental or due to chance as much as anything. The question, as clarified by Hedges (1994), is not "What is true about these studies?" but "What is true about studies like these that could have been done?" Such generalizations can be handled through statistical means by incorporating the additional uncertainty due to the inference to studies not identical to those in the sample. Upon identification of the appropriate universe of generalization, the synthesis can proceed. However, there are additional considerations.

Hedges (1994) also suggested that all of the available methods commonly used in statistical analysis of study results assume two conditions. The statistic or estimated effect (or appropriate transformation) is normally distributed in large-samples with a mean that approximates the parameter of interest. In addition, the standard error of the estimated effect is a

continuous function of its study sample size, the magnitude of the effect, and potentially other factors that can be estimated from the resulting study data. Once these considerations have been addressed, results from studies can be combined.

If studies were actually identical replications of an original study, combining results could be a relatively straightforward exercise. In reality, studies differ in many ways, some of which were described above, but also in terms of location, timing, and sample size. To deal with these differences, the synthesist has available a number of weighting schemes based on three assumptions:

> (a) Theory or evidence suggests that studies with some characteristics are more accurate or less biased with respect to the desired inference than studies with other characteristics, (b) the nature and direction of that bias can be estimated prior to combining, and (c) appropriate weights to compensate for the bias can be constructed and justified. (Shadish & Haddock, 1994, p. 263)

Based on the notation used in the *Handbook of Research Synthesis* (see Raudenbush, 1994), in a random effects model, θ_i (the population parameter of interest) is not fixed, but random with its own distribution. Total variability of an observed study estimate includes both conditional variation, v_i, of the estimate around each population θ_i and random variation σ_θ^2 of the θ_i around the mean population parameter. Estimation of the random variance component (the between-studies variance) is no easy matter. Various estimation procedures result in different estimates with important consequences (see Raudenbush, 1994). In this context, the weights, w_i, that minimize the variance of the estimated mean population parameter $\overline{T}_\bullet$ are inversely proportional to the variance, where $w_i = \frac{1}{v_i + \sigma_\theta^2} = \frac{1}{v_i^*}$. Uncertainty, when considering sample studies to be representative of a larger universe, comes from the fact that

study contexts, treatments, and administration procedures differ in many ways that potentially impact results.

Threats to valid inferences from such a synthesis remain. Among those described by Raudenbush (1994) are uncertainty about the random effects variance component, the tenability of the assumption that the random effects are normally distributed with constant variance, model misspecification, and multiple effects from single studies resulting in dependent data. None of the reliability generalization studies reviewed above addressed these issues.

An IRT Perspective

Item response theory (IRT) is a general statistical theory that relates item and test performance to abilities or traits measured by the items in a test. A common mathematical form of this relationship is a logistic model that links item performance to an unobservable ability or trait level. $P_i(\theta)$ is the probability of a correct response to item ***i*** as a function of ability (θ, theta). In a more general sense, in noncognitive measures, this corresponds to the probability of a certain item response given the individual's trait level. The resulting graph of the relationship between probability of an item response and trait level is the item characteristic curve (ICC). The sum of the ICCs for a given test constitute the test characteristic function, used to predict scores at given trait levels. So for a given set of items, an individual's expected score at a given trait level is their true score. When the mathematical model expressing these probabilities fits the data, the item and person parameters are invariant or sample independent.

A useful contribution of IRT is the specification of the item information function, which contains the value of the item to the assessment of the trait. The test information function, $I(\theta)$, is the sum of each item information function for a given instrument and provides an estimate of

the trait estimation error. A standard error can be obtained for each estimated trait level, $\hat{\theta}$, where $SE(\hat{\theta}) = \sqrt{\frac{1}{I(\hat{\theta})}}$. From this we can see that smaller errors of estimation result from those trait levels where the test provides the most information. For a more thorough comparison of IRT and classical test theory, see Hambleton and Jones (1992).

The IRT analogue to the classical test theory standard error can be conceived of as the error variance for a given theta (trait level) integrated over the theta distribution: $\sigma_e^2 = \int \sigma_{\varepsilon|\theta}^2 \, g(\theta) d\theta$, where $\sigma_{\varepsilon|\theta}^2 = \frac{1}{I(\theta)}$. The information function is fixed by the items from the instrument; it is the sum of each item information function for a given instrument. So to obtain a different estimate of the standard error, we would need a different theta distribution—a sample with a different trait distribution.

In this framework, changes in σ_e^2 (error variance) or reliability is indicative of a change in the theta distribution. In IRT, error variance or reliability is based on the fixed calibration settings so that the error term is fixed given the items on the test, the setting, etc.

Citing Lord and Novick (1968), Samejima argued that

> unlike classical test theory, latent trait theory provides us with the standard error of estimation as a measure independent of any group of examinees, and given locally, or as a function of ability θ. For this reason, we can consider this an intrinsic property of the test itself, as long as the populations of our interest belong to the complete latent space. (p. 237-238)

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington DC: AERA.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60*(2), 236-254.

Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (eds.), *The handbook of research synthesis* (pp. 3-14). New York: Russell Sage Foundation.

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper & L. V. Hedges (eds.), *The handbook of research synthesis* (pp. 17-28). New York: Russell Sage Foundation.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (eds.), *The handbook of research synthesis* (pp. 29-38). New York: Russell Sage Foundation.

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

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*(2), 233-247.

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*(2), 157-173.

Sawilowsky, S. S. (2000b). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement, 60*(2), 196-2000.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183-198.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174-195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20.

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*(2), 224-235.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*(2), 201-223.