The present article addresses reliability issues in light of recent studies and debates focused on psychometrics versus datametrics terminology and reliability generalization (RG) introduced by Vacha-Haase. The purpose here was not to moderate arguments presented in these debates but to discuss multiple perspectives on score reliability and how they may affect research practice, editorial policies, and RG across studies. Issues of classical error variance and reliability are discussed across models of classical test theory, generalizability theory, and item response theory. Potential problems with RG across studies are discussed in relation to different types of reliability, different test forms, different number of items, misspecifications, and confounding independent variables in a single RG analysis.

The editorial policies of Educational and Psychological Measurement (EPM) have become a focal point of recent studies and debates related to important issues of reliability, effect size, and confidence intervals in reporting research. The purpose of this article was not to reiterate or moderate datametrics-psychometrics arguments presented in these debates (e.g., Thompson, 1994; Sawilowski, 2000; Thompson & Vacha-Haase, 2000) but to discuss issues concerning the assessment of score reliability and how they may affect (current or future) research practice, editorial policies, and reliability generalization (RG) across studies introduced by Vacha-Haase (1998). The discussion is organized in three parts. The first part presents multiple perspectives on classical standard error of measurement (SEM) and reli-

I would like to thank the two anonymous reviewers and an acting editor for their helpful comments and suggestions. Correspondence regarding this article should be addressed to Dimiter M. Dimitrov, Ph.D., 507 White Hall, College of Education, Kent State University, Kent, OH 44242-0001.

Educational and Psychological Measurement, Vol. 62, No. 5, October 2002 783-801
DOI: 10.1177/001316402236878
© 2002 Sage Publications
ability across traditional and modern measurement models. The second part addresses potential problems with RG across studies in light of issues discussed in the first part. The conclusion summarizes major points of the discussion in the first two parts in an attempt to provide some insight for the practice of substantive research and RG across studies.

In response to *EPM* editorial policies stating that “scores, not tests, are reliable” (Thompson, 1994; Thompson & Daniel, 1996), Sawilowski (2000) addressed problems inherent with datametrics from the perspectives of an historical review of reliability terminology and textbook treatments of reliability. He recognized that statements related to “reliability of the test” are not sufficiently informative because “reliability paradigms and their coefficients are simply not interchangeable” (p. 159). He indicated, however, that reliability estimates should not be tied only to the data at hand:

Indeed, the purpose for using a nationally representative scientifically selected random sample when conducting reliability studies during the test construction process is to obtain the reliability estimate of a test for general purposes. In my view, authors ought always to report these reliability coefficients from manuals and other sources along with the reliability estimate obtained from, and a description of, the researcher’s own sample. (p. 170)

This point of view, based on numerous treatments of reliability (e.g., Crocker & Algina, 1986; Goodenough, 1949; Gronlund, 1988; Nunnally & Bernstein, 1994; Payne, 1975; Suen, 1990), is also represented by Traub (1994) who noted that “the usual reliability experiments provide a sample estimate of the reliability coefficient for the population” (p. 66). On their side, Thompson and Vacha-Haase (2000) recognized that the treatment of reliability should not be isolated from the context of the measurement theory because different ways of estimating reliability involve different sources of error. The author of this article, therefore, argues that the understanding of reliability and related policies for calculating and reporting reliability should be based on multiple perspectives about population reliability and its sample estimates. Inferences from such perspectives can also be used to identify potential problems with RG across studies.

Reliability in Classical Test Theory

*Cronbach’s α and Reliability of Congeneric Measures*

Under classical test theory, it is important to make distinctions between models that are parallel, essentially τ equivalent, or congeneric (e.g., Crocker & Algina, 1986; Feldt & Brennan, 1989; Lord & Novick, 1968; Raykov, 1997).
Thompson and Vacha-Haase (2000) referred to these models and the possibility of testing their assumptions using structural equation modeling (Jöreskog & Sörbom, 1989), but the EPM guidelines editorial of Fan and Thompson (2001) did not elaborate on this issue in discussing confidence intervals for reliability estimates. Therefore, some specific comments seem appropriate.

Let us have the test divided into $k$ components; $2 \leq k \leq n$, where $n$ is the number of items in the test. By assumption, the score $X_i$ on each component can be represented as $X_i = T_i + E_i$, where $T_i$ and $E_i$ denote the true score and error score, respectively. The test components are called parallel if any two items $i$ and $j$ have equal true scores, $T_i = T_j$, and equal error variances, $\text{Var}(E_i) = \text{Var}(E_j)$. With the assumption of error variances dropped, the test components are called (a) essentially tau-equivalent, if $T_i = T_j + a_{ij}$, where $a_{ij} \neq 0$ is a constant; and (b) congeneric if $T_i = b_{ij}T_j + a_{ij}$, where $b_{ij}$ is a constant ($b_{ij} \geq 0$, $b_{ij} \leq 1$).

The reliability of the test for a population, $\rho_{xx}$, is defined as the ratio of true score variance to observed score variance, $\text{Var}(T_i)/\text{Var}(X_i)$, or (equivalently) as the squared correlation between true and observed scores (e.g., Lord & Novick, 1968, p. 57). In empirical research, however, true scores cannot be directly determined and, therefore, the reliability is typically estimated by coefficients of internal consistency, test-retest, alternate forms, and other reliability methods adopted in the psychometric literature (e.g., Crocker & Algina, 1986). One of the most frequently used method of estimating internal consistency reliability is Cronbach’s coefficient $\alpha$ (Cronbach, 1951). It is important to emphasize that confidence intervals for Cronbach’s $\alpha$ (or other reliability estimates) should also be calculated and reported for the data in hand with substantive studies. Classical approaches for determining confidence intervals about score reliability coefficients are discussed, for example, in an EPM editorial (Fan & Thompson, 2001).

It should be noted, however, that (even at population level) Cronbach’s $\alpha$ is an accurate estimate of reliability, $\rho_{xx}$, only if there is no correlation among errors and the test components are at least essentially tau-equivalent (Novick & Lewis, 1967). As a reminder, tau equivalency implies that the test components measure the same trait and their true scores have equal variances in the population of respondents. When errors do not correlate but the test components are not essentially tau-equivalent, Cronbach’s $\alpha$ will underestimate $\rho_{xx}$. If, however, there is a correlation among error scores, Cronbach’s $\alpha$ may substantially overestimate $\rho_{xx}$ (see Komaroff, 1997; Raykov, 2001; Zimmerman, Zumbo, & Lalonde, 1993). Correlated error may occur, for example, with adjacent items in a multicomponent instrument, with items related to a common stimulus (e.g., same paragraph or graph), or with tests presented in a speeded fashion (Komaroff, 1997; Raykov, 2001; Williams & Zimmerman, 1996).
Evidently, it is important that EPM editorial policies address this issue and that related discussions elaborate on congeneric measures when assumptions of parallel or essentially tau-equivalent test components do not hold. Moreover, previous research provides sound theoretical and practical treatments of reliability for congeneric measures. For example, Raykov (2001, 1997) proposed structural equation modeling methods for obtaining confidence intervals of reliability with congeneric tests as well as computer applications of these methods with EQS (Bentler, 1995). Relatively simple procedures for determining weights that maximize reliability under a congeneric model are provided, for example, by Wang (1998). As a reminder, congeneric tests are measures of the same latent trait, but they may have different scale origins and units of measurement, and may vary in precision (Jöreskog, 1971). Thus, congeneric measures facilitate aggregation or comparison of reliability coefficients across different scales and forms of measurement instruments. This can be very beneficial for RG across studies that use and report congeneric scales. Although congeneric measures were not widely used in previously published studies, investigators and editorial policies should target applications of such measures in future research.

Test-Retest Reliability

Test-retest reliability is estimated by the correlation between the observed scores of the same people taking the same test (or parallel test forms) twice. However, this procedure may run into problems such as carry-over effects due to memory and/or practice (Cronbach & Furby, 1970; Lord & Noivick, 1968). Test-retest reliability estimates are most appropriate for measuring traits that are stable across time period between the two test administrations (e.g., work values, visual or auditory acuity, and personality). It is important to note that test-retest reliability and internal consistency reliability are independent concepts. Basically, they are affected by different sources of error, and therefore, it may happen that measures with low internal consistency have high temporal stability and vice versa (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994). Previous research on stability in reliability showed that the test-retest correlation coefficient can serve reasonably well as a surrogate for the classical reliability coefficient if an essentially tau-equivalent model with equal error variances or a parallel model is present (Tisak & Tisak, 1996). In fact, recent RG studies also report substantial reliability variation due to type of reliability coefficients (e.g., Vacha-Haase, 1998; Yin & Fan, 2000).

Alternate Form Reliability

Alternate form reliability is a measure of the consistency of scores across comparable forms of a test administered to the same group of individuals
DIMITROV 787

(Crocker & Algina, 1986). Ideally, the alternate forms should be parallel, but this is difficult to achieve, especially with personality measures. The correlation between observed scores on two alternate test forms is usually referred to as the coefficient of equivalence. Basically, the alternate form reliability and internal consistency reliability are affected by different sources of error. If the correlation between alternate forms is much lower than the internal consistency coefficient (e.g., by 0.20 or more), this might be due to (a) differences in content, (b) subjectivity of scoring, and (c) changes in the trait being measured over the time between the two administrations of alternate forms. To determine the relative contribution of these sources of error, it is recommendable to administer the two alternate forms (a) on the same day for some respondents and (b) within a 2-week time interval for others (see, Nunnally & Bernstein, 1994; Dimitrov, Rumrill, Fitzgerald, & Hennessey, 2001). The comments in this section are also important for the discussion of RG across studies later in this article.

Reliability in Generalizability Theory

Generalizability (G) theory simultaneously takes into account all available error sources (facets) such as items, raters, test forms, and occasions that influence the reliability for either relative or absolute (criterion-related) decisions (Cronbach, Gleser, Nanda, & Rajaratnum, 1972). Classical test theory takes into account only one facet (e.g., items) at a time and provides estimates of reliability only for relative decisions. The one-facet crossed design “persons x items” (p x i) in G theory corresponds to the classical model when persons respond to items. With this design, the raw-score variance, $\sigma^2_{xt}$, is partitioned into three independent variance components: (a) $\sigma^2_i$, which indicates how much items differ in difficulty; (b) $\sigma^2_p$, which indicates how much persons differ in their responses; and (c) $\sigma^2_{pi,e}$, for the confounding effect of the $p \times i$ interaction with other sources of variation, e. The error variance for relative decisions is $\sigma^2_{Rel} = \sigma^2_{pi,e} / n$ and the error variance for absolute decisions is $\sigma^2_{Abs} = (\sigma^2_i + \sigma^2_{pi,e}) / n$, where $n$ is the number of items. With many-facet designs, $\sigma^2_{Rel}$ is defined as all variance components that affect the relative standing of persons and $\sigma^2_{Abs}$ as all variance components (except for the object of measurement, $\sigma^2_p$) that contribute to measurement error. In G theory, the reliability coefficients for relative decisions (generalizability coefficient, $\rho^2_{Rel}$) and absolute decisions (index of dependability, $\Phi$) are defined as follows (e.g., Shavelson & Webb, 1991, p. 93):

$$\rho^2_{Rel} = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{Rel})} \quad \text{and} \quad \Phi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{Abs})} \quad (1)$$

Using the results from a G study, one can also control SEM and reliability levels by manipulating the conditions of each facet in a decision study. Previous
research provided numerous theoretical, methodological, and practical treatments of reliability in G theory (e.g., Brennan, 1983; Clauser, Clyman, & Swanson, 1999; Crick & Brennan, 1983; Erlich & Shavelson, 1978; Sun, Valiga, & Gao, 1997; Shavelson & Webb, 1991). Some advantages of the G theory reliability coefficients are also discussed by Thompson and Vacha-Haase (2000).

**Item Response Theory Perspective**

**on Classical SEM and Reliability**

In item response theory (IRT), the term *ability* (or *trait*, for psychological constructs) connotes a latent trait that underlies the responses of the participants on items of the measurement instrument. The ability score, $\theta$, of an examinee on an achievement test determines the probability for this examinee to answer correctly any test item (Lord, 1980). Although the classical test theory is concerned with the accuracy of observed scores, the IRT is concerned with the accuracy of ability scores. In IRT, the conditional error variance at an ability level $\theta$ is inversely related to the amount of information, $I(\theta)$, provided by a test at $\theta$ (e.g., Lord, 1980; Samejima, 1977). This section, however, does not deal with accuracy of $\theta$ but, instead, with classical SEM and reliability from IRT perspective.

Lord (1957) noted that “a reliability coefficient depends on the spread of ability among the examinees, as well as the number of items in the test, the discrimination power of the items, and other factors” (p. 510). Lord (1980) related classical SEM and reliability coefficients to IRT estimates of item parameters and ability. He presented the error variance of number-right scores on a test of $n$ dichotomous items as the mean of conditional error variances at the ability levels of $N$ examinees, $\theta_1, ..., \theta_n$ (Lord, 1980, p. 52):

$$
\hat{\sigma}_e^2 = \frac{1}{N} \sum_{j=1}^{N} \sigma_{x_i j}^2 = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{n} P_i(\theta_j)[1 - P_i(\theta_j)]
$$

In Equation 2, $P_i(\theta_j)$, the probability for correct answer on item $i$ from an examinee with ability $\theta_j$, is calculated with the respective IRT model. The product $P_i(\theta_j)[1 - P_i(\theta_j)]$ is the conditional error variance for the dichotomous item $i$ at ability level $\theta_j$. The square root of the error variance in Equation 2 is a classical (sample-based) SEM. The population error variance (expected value of the sample error variance) is obtained by replacing the summation over discrete ability scores, $\theta_1, ..., \theta_n$, with integration over the ability continuum:

$$
\sigma_e^2 = \sum_{i=1}^{n} \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)] \phi(\theta) d\theta,
$$

(3)
where $\varphi(\theta)$ is the probability density function (pdf) for the ability distribution. The integration is from $-4$ to $4$ because the ability, $\theta$, is not limited in the theoretical framework of IRT. The ability scores typically vary from $-4$ to $4$ on the logit scale (e.g., Lord, 1980).

An important property of Equation 3 is that the integral components in the summation of $n$ items represent additive yet independent contributions of the individual items to the classical error variance. This feature, not available with other methods of estimating classical SEM and error variance, may be very useful for the practice of test development and reliability analysis. For example, given the IRT calibration of items (available with most standardized instruments), one can select items to develop a measurement instrument with prespecified error variance. Then, one can also determine the classical SEM and reliability, $\rho_{xx}$, from the relationships:

$$SEM = \sqrt{\sigma_e^2} \quad \text{and} \quad \rho_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$ (4)

May and Nicewander (1993) used Equation 3 in comparing reliability for number-right scores and percentile ranks. Dimitrov (2001a, 2001b) provided exact formulas for the classical error variance in Equation 3 under different distributions (e.g., normal, triangular, and logistic) of the ability (trait) scores for a population of participants.

From Rasch measurement perspectives, Wright (2001) also noted that “once the test items are calibrated, the standard error corresponding to every possible raw score can be estimated without further data collection” (p. 786). For a 14-item test, for example, he found that the “sample-independent” reliability for the test (with a separation index of 4) was .94, whereas the empirical reliability was .62, thus “grossly underestimating the test’s measurement effectiveness” (p. 786).

Evidently, classical SEM and reliability can be estimated from IRT information about item parameters and trait distribution for the population. Such information is available (or easy to obtain) with standardized measurement instruments. This perspective on SEM and reliability has important implications for understanding, estimating, and comparing accuracy of measurements. For example, the error variance in Equation 3 is a classical match of the marginal error variance that is used in computer adaptive testing for comparison between internal consistency reliability of a computer adaptive test and alternatively used paper-and-pencil forms (e.g., Thissen, 1990).

Additional Perspectives on Reliability

Other types of reliability, such as interrater reliability, criterion-referenced reliability (classification consistency), and stability in reliability for repeated measurements, can also be treated from different perspectives. In previous
research, interrater reliability has been estimated with correlations (e.g., Fan & Thompson, 2001), factorial design (Winer, 1971), Cohen’s kappa (Cohen, 1960), extended (g-wise) kappas (Conger, 1980), G theory coefficients (e.g., Brennan & Prediger, 1981; Cooil & Rust, 1994), latent class models (e.g., Dillon & Mulani, 1984), and Rasch measurement models (e.g., Linacre, 1994). Criterion-referenced reliability has been also estimated through classical statistics (e.g., Berk, 1980; Huynh, 1976; Peng & Subkoviak, 1980), G theory (Brennan & Kane, 1977), IRT (e.g., Hambleton & Novick, 1973), and latent class models (e.g., Van Der Linden & Mellenbergh, 1977). Stability in reliability for repeated studies has been evaluated primarily through structural equation modeling (e.g., Jöreskog, 1971; Raykov, 2000; Tisak & Tisak, 1996). It should be noted that stability in reliability for repeated studies uses the same test and same examinees across time points, whereas the RG across studies involves different samples and may use (which is questioned later in this article) different types of reliability, different test forms, and different number of items in a single analysis.

Reliability of measurement of change is another perspective on reliability related to pre- and posttest scores. This type of reliability has been discussed extensively throughout the history of psychometrics (e.g., Collins, 1996; Cronbach & Furby, 1970; Kane, 1996; Linn & Slindle, 1977; Rogosa, 1995; Zimmerman & Williams, 1982). The common misconception that the gain scores always have low reliability and thus have questionable value still affects the way most substantive studies report and interpret their results. Zimmerman and Williams (1982) demonstrated that the reliability of gain scores can be consistently high when the pre- and posttest measures have unequal standard deviations and unequal reliability coefficients. Recently, Kane (1996) showed that the reliability of gain scores can be interpreted as a context-specific, between-examinees precision index.

**SEM and Reliability**

Previous studies indicate that although the reliability coefficient is a convenient unitless number between 0 and 1, the SEM relates to the meaning of the scale and is, therefore, more useful for score interpretations (e.g., Feldt & Brennan, 1989; Gronlund, 1988; Thissen, 1990). The classical relationship between reliability and SEM in equation (4) is based on the (generally false) assumption that the error variance is the same for all scores (e.g., Thissen, 1990; Thompson, 1992; Thorndike, 1982). The accuracy of measurement is therefore much better represented by conditional SEM estimates at different score levels. Procedures for obtaining such estimates with the classical theory are provided in previous research (e.g., Kolen, Hanson, & Brennan, 1992; Thorndike, 1982). Kolen at al. (1992) noted,
Estimation of conditional standard errors of measurement of scale scores is important, especially because the Standards for Educational and Psychological Testing (AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education], 1985), particularly Standard 2.10, recommends that conditional standard errors be reported. In addition, conditional standard errors of measurement are valuable to users of test results for constructing confidence intervals for reported scores and for communicating an index of measurement error. (p. 286)

Traditionally, the classical error variance is estimated as

\[ \sigma^2_e = \sigma^2_x (1 - \rho_{xx}), \]  

where \( \rho_{xx} \) is substituted for a sample-based estimate of reliability (e.g., Cronbach’s \( \alpha \)). Whereas Equation 5 is technically appropriate for the data in hand with substantive studies, formula (3) provides important population perspectives on classical error variance. One can see, for example, that (a) the error variance does not depend on reliability estimates, and (b) given the shape of the trait distribution, the independent additive contribution of individual items to the error variance can be determined from their parameters (e.g., difficulty and discrimination). This perspective on classical error variance is consistent with empirical findings that SEM remains fairly constant across samples, but this is not true for the reliability coefficient, which is dependent on the score variance in the sample tested (see, Gronlund and Linn, 1990, p. 92; Sawilowski, 2000, p. 164).

Potential Problems
With RG Across Studies

The comments on RG across studies in this section relate primarily to meta-analytic tools used by Vacha-Haase (1998) and do not generalize over all possible ways of conducting RG. In fact, in response to the criticism of Sawilowski (2000) to some of the analytic choices made by Vacha-Haase (1998), Thompson and Vacha-Haase (2000) noted, “We do not see RG as invoking a monolithic methodological approach” (p. 187) and, “We do not see RG as involving always a single genre of analyses” (p. 185). Although I totally agree with this, there are several points to be made about the RG methodology introduced by Vacha-Haase (1998) and replicated in numerous RG studies recently published in EPM (e.g., Capraro, Capraro, & Henson, 2001; Caruso, 2000; Caruso, Witkiewitz, Belcourt-Dittloff, & Gottlieb, 2001; Henson, Kogan & Vacha-Haase, 2001; Vacha-Haase, Kogan, Tani, & Woodall, 2001; Viswesvaran, & Ones, 2000; Yin & Fan, 2000). Most argu-
ments in this section are based on facts and comments provided in previous sections.

Using Different Types of Reliability

Using different types of reliability in a single RG analysis may lead to mixing apples and oranges, thus violating basic meta-analytic principles (e.g., Hunter & Schmidt, 1990). For example, Vacha-Haase (1998) used multiple regression with the dependent variable being either test-retest correlation coefficients or Cronbach’s $\alpha$ (or KR-20) coefficients. However, none of these coefficients will represent the classical reliability unless specific assumptions can be made. As a reminder, Cronbach’s $\alpha$ can either underestimate the reliability when the measures are not at least essentially tau-equivalent (Novick & Lewis, 1967) or overestimate the reliability, when correlations among errors occur (Komaroff, 1997). Also, the test-retest correlation coefficient can be a reasonable estimate of reliability only if the measures are essentially tau-equivalent and have equal error variances (Tisak & Tisak, 1996). Thus, the dependent variable used by Vacha-Haase (1998) might be a mixture of apples and oranges, as implied also by her own conclusion that “the results ... indicate that internal consistency and test-retest reliability coefficients seem to present considerably different pictures of score quality” (p. 16). Yin and Fan (2000) also noted that

the results for partitioning the variances of the BDI [Beck Depression Inventory] score reliability estimates indicated that the type of reliability coefficients (internal consistency vs. test-retest reliability coefficients) and type of study participants ... were statistically significant and practically meaningful predictors in the general linear model with reliability estimates as the dependent variable. (p. 216)

Given that most published studies do not test for tau-equivalency or other assumptions related to reliability coefficients they report (when they do), RG researchers should be cautious about mixing internal consistency and test-retest reliability coefficients in a single RG analysis.

Using Tests With Different Lengths

A problem with RG across studies may also result from using alpha coefficients for tests with different lengths. Some RG investigators use the Spearman-Brown formula to adjust alpha coefficients (e.g., Caruso, 2000). It should be noted, however, that the Spearman-Brown formula produces $\alpha$ coefficients only under the assumption of equal item variances (Charter, 2001). Although it is clear that RG researchers do not have access to raw item-level data and thus cannot test for equal item variances, it is recom-
mended that they reflect this in the discussion of their results. Other RG across studies do not adjust the $\alpha$ coefficients and instead use dummy coding for the number of items administered (e.g., Vacha-Haase, 1998; Viswesvaran & Ones, 2000; Yin & Fan, 2000). Let us also remember that the split-half approach of estimating internal consistency requires that the two parts of the test have equal variances (Kristof, 1970). One can also refer to Thorndike (1982, p. 148) for a statistical analysis of effects of manipulating test length on true-score variance, error variance, and reliability, as well as to Alsawalme and Feldt (1999), for statistical tests for hypotheses about alphas extrapolated from the Spearman-Brown formula. Awareness of such effects can also help RG researchers to adequately interpret and discuss RG study results.

Using Tests With Different Formats

Using different test forms in a single RG analysis can also cause problems because, as indicated by previous research, the reliability depends on factors such as item response format (Frisbie & Druva, 1986), (positively/negatively) wording of stems (Barnette, 2000), and type of scales (Cook, Heath, Thompson, & Thompson, 2001). For example, estimating the reliability of multiple true-false (MTF) tests, Frisbie and Druva (1986) noted,

If the MTF items in the same cluster [items sharing a common stem] are more highly intercorrelated than are MTF items that appear in different clusters, then some conventional internal analysis procedures for estimating test reliability may yield spurious results. The results may be spurious because the assumptions required by many of the procedures for estimating reliability include subtest equivalence or parallel items. The presence of systematic patterns of intercorrelations could violate either assumption. (p. 100)

Regarding the role of stem wording, Barnette (2000) noted that “negated items are not considered the exact opposite of directly worded items, and this is one of the major factors in the reduction of the reliability and validity of scores on surveys using mixed items” (p. 369). Cook et al. (2001) studied the dependence of reliability on the type of scale in Web- or Internet-based surveys, noting that

although greater score variance is possible when more score intervals are employed, if participants do not cognitively process so many scores intervals when dealing with unnumbered graphic scales [a continuous line drawn between two antonyms], then the use of excessive intervals may not improve score reliability and might even lessen reliability. ( p. 705)

Evidently, different test forms should be used with proper understanding and awareness of problems that they may cause in a single RG analysis.
Confounding Independent Variables

Another problem with RG across studies may result from inappropriate coding of groups (e.g., by gender, ethnicity, age, status) or, even worse, from biasedness of the instrument on groups coded with independent variables. For example, Sawilowski (2000) noted that the coding design for gender in the RG study of Vacha-Haase (1998) led to a confounding of independent variables since gender of study participants was coded twice. First, gender was coded 1 when the study participants were both men and women and 0 when they were all men or all women. Then, gender was coded 1 when the study participants were all women and 0 when they were all men or both men and females. The negative effect of confounding independent variables such as gender and ethnicity will increase if the instrument is biased against a specific group in the context of some substantive studies used with the RG analysis. As a reminder, an instrument will be, for example, gender biased if male and female participants who share the same location on the trait being measured by the instrument consistently differ in their responses. Evidently, this will obscure the actual contribution of gender to the reliability variation across RG studies. Bias effects may occur in RG studies when, for example, the instrument is used in different contexts of age, language, and cultural experience (e.g., Yin & Fan, 2000).

Sample-Related Misspecifications

Problems with RG may also result from misspecifications that occur when relevant characteristics of the study samples are not coded as independent variables in RG analysis. For example, Sawilowski (2000) noted that failure to consider random samples versus nonrandom samples led to misspecifications in the RG regression analysis of Vacha-Haase (1998). Another source of possible misspecifications, not addressed in previous RG studies or debates, relates to failure to take into account the underlying distribution of scores for the study samples. Yet, Lord (1984) noted that “unless the distribution of ability is known, it is not logically possible to get an unbiased estimate of the SEM for examinees who are chosen for consideration because they all have a given observed score” (p. 241). As one can also see from Equation 3, the classical error variance is affected by the distribution of the trait underlying the participants’ responses. Thus, composition factors that may affect the trait distribution of study samples are particularly relevant for the practice of RG. Information about such composition factors, if not provided with studies used in RG, may come from previous theoretical or empirical findings. Avoiding misspecifications related to group homogeneity or differences in trait distributions for study samples (or composition subgroups) will reduce the risk of capitalization on chance in RG.
**RG Across Studies Versus Accuracy of Measurement Within Studies**

The results from RG across studies may provide potentially useful information about the variability of reliability coefficients, but they do not represent the accuracy of measurement for a specific study. Indeed, the accuracy of measurement for a study relates to information about the (study-specific) population reliability and its sample estimates. Within the classical framework, such information is provided, for example, by confidence intervals for alphas (Fan & Thompson, 2001; Pedney & Hubert, 1975) or reliability for congeneric measures (Raykov, 1998), but not by reliability box-plots in RG across studies (e.g., Vacha-Haase, 1998, p. 13). When IRT calibration of items is available, classical SEM can be determined with Equation 3 (e.g., Dimitrov, 2001a, 2001b).

It should be noted that although the unitless reliability coefficient is more useful for comparing the accuracy of measurement across scoring procedures, the SEM is more useful with a specific study because it provides scale-specific interpretations of the accuracy of measurement for this study. However, generalization of SEM across studies cannot be done with RG analyses. As Thompson and Vacha-Haase (2000) noted,

> An RG analysis of SEMs across studies would not be reasonable when the meta-analysis looked at variations in score quality in range of measures ..., nor would an SEM focus be appropriate even when a single measure was investigated if both short and long forms were used in the prior studies. (p. 188)

It is difficult to understand, then, why some RG studies (e.g., Yin and Fan, 2000) include SEM (in addition to reliability) in a single RG analysis with different forms of the measurement instrument and different types of reliability coefficients. Moreover, the findings about SEMs in such RG studies do not relate to interpretational characteristics of SEMs, but to theoretically based statistical artifacts. For example, Yin and Fan (2000) reported that “SEMs were not correlated with reliability estimates, but SEMs were substantially related to the standard deviations of the BDI scores in different studies” (p. 217).

**Conclusion**

The multiple perspectives on SEM and reliability addressed in this article have important implications for (a) understanding, estimating, and reporting accuracy of scores within studies and (b) identifying potential problems with the RG across studies. Clearly arguments could be made for improving current and future practices of estimating and reporting SEM and reliability in
substantive or measurement studies. The facts and comments provided in this article lend support to recent editorial policies of *EPM* about reporting confidence intervals for score reliability coefficients (Fan & Thompson, 2001). However, researchers should also be encouraged to calculate and report more appropriate estimates of SEM and reliability in light of multiple perspectives on these two concepts in traditional and modern measurement models. For example, assumptions for the traditionally used Cronbach’s coefficient $\alpha$ should be checked and, if necessary, researchers should report reliability for congeneric measures. Subsequent RG across studies will also benefit from this because congeneric tests relate to the same trait, but they may vary in precision, scale origins, and units of measurement (Jöreskog, 1971). It is recommended also that researchers use, whenever possible, generalizability theory methods to take into account all available error sources (e.g., items, raters, and occasions) that influence the reliability for either relative or absolute decisions.

It should be emphasized that the IRT section in this article presents an IRT perspective on classical SEM and reliability, not the IRT treatment of conditional standard errors at trait levels. It was shown that, given the shape of the trait distribution, the independent additive contribution of individual items to the classical SEM depends only on IRT estimates of their parameters (e.g., difficulty and discrimination). Thus, when such estimates are available (e.g., with standardized measurement instruments), researchers may estimate classical SEM for the study population with higher quality and accuracy relative to traditional approaches (e.g., confidence intervals about Cronbach’s $\alpha$). This perspective on SEM and reliability also shows that “population reliability of the test” connotes a sound psychometric concept. Therefore, although agreeing that “the test qua test is not reliable or unreliable” (Thompson & Vacha-Haase, 2000), we should also agree that concepts related to accuracy of measurement are psychometric concepts that can be treated from multiple perspectives within and across theoretical frameworks. From terminology point of view, then, the necessity of the term *datametrics* is questionable if its sole purpose is to convey that “data, not tests, are reliable” in psychometrics (e.g., Thompson & Vacha-Haase, 2000). Also, this term is loosely defined and semantically problematic because data result from measurement, we do not measure data. Conversely, psychometrics is well defined as “related to measurement of attributes of people or other objects, with tests that are intended to measure those attributes, and with single exercises or items that in combination make such tests” (Thorndike, 1982, p. 4).

The potential problems with RG across studies identified in this article relate primarily to (a) using different types of reliability, different number of items, and/or different test forms in a single RG analysis; (b) confounding independent variables; and (c) misspecifications related to group homogene-
ity, test bias, and irrelevance of composition factors. It should be emphasized that RG researchers were not able to address most of these problems because prior published studies reported very limited information about the accuracy of their data (in the rare cases that they reported such information at all). What RG researchers should do, however, is discuss this issue in light of possible pitfalls and limitations of their RG procedures and results. Disclaimers for RG articles to include in their discussion relate, for example, to using Cronbach’s α, although (a) the test components are likely congeneric, suggesting that alpha may underestimate the reliability; (b) there are chances for correlated errors (e.g., items sharing a common stem), suggesting that Cronbach’s α may overestimate the reliability; and (c) the tests are of different length, suggesting possible loss of accuracy in generalizing alpha; (it is recommended that an adjustment for alpha, using the Spearman-Brown formula, be performed, with a word of caution when the assumption of equal item variances with this formula is not checked).

Other disclaimers in RG discussions may relate to using different types of reliability in a single RG analysis, which may produce variability effects due to statistical artifacts (e.g., Hunter & Schmidt, 1990). Also, using different test forms may produce spurious effects and reduce the reliability due to (a) systematic intercorrelations among items sharing a common stem in multiple true-false test, (b) negative stem wording, or (c) excessive scale intervals. Other possible sources of disclaimers in RG discussions relate to the following: (a) confounding independent variables; (b) presence of bias when the instrument is used in different contexts of age, language, or cultural experience; and (c) misspecifications due to failure to take into account sample characteristics (e.g., randomness or shape of the trait distribution) that may affect the reliability coefficients.

It should be also noted that RG across studies does not provide adequate information about the accuracy of measurement with a specific study. Such information is provided, instead, by SEM and reliability coefficients for the study sample and (exact or interval) estimates of their population parameters (e.g., Dimitrov, 2001a, 2001b; Fan & Thompson, 2001). Nor can the RG across studies assess stability (time-invariance) in reliability. This can be done, instead, with longitudinal models for generalization of reliability using the same instrument with the same sample across time points (e.g., Tisak & Tisak, 1996).

In conclusion, with awareness of potential problems and appropriate designs, RG across studies can provide useful meta-analytic information about variability of reliability coefficients. I believe, however, that the methodological soundness and practical usefulness of RG studies depend, among other things, upon implications from the multiple perspectives on reliability presented in this article.
References


