Organizational Research Methods
Volume 12 Number 1
January 2009 165-200

10.1177/1094428107302900
http://orm.sagepub.com
hosted at
http://online.sagepub.com

The Multifaceted Nature of Measurement Artifacts and Its Implications for Estimating Construct-Level Relationships

Huy Le
University of Central Florida
Frank L. Schmidt
University of Iowa
Dan J. Putka
Human Resources Research Organization

Measurement artifacts, including measurement errors and scale-specific factors, distort observed correlations between measures of psychological and organizational constructs. The authors discuss two alternative procedures, one using the generalized coefficient of equivalence and stability (GCES) and one based on structural equation modeling, to correct for the biasing effect of measurement artifacts in order to estimate construct-level relationships. Assumptions underlying the procedures are discussed and the degrees of biases resulting from violating the assumptions are examined by means of Monte Carlo simulation. They then propose an approach using cumulative knowledge in the literature about properties of measures of a construct to estimate the GCES. That approach can allow researchers to estimate relationships between constructs in most research situations. The authors apply the approach to estimate the GCES for overall job satisfaction, an important organizational construct.

Keywords: *measurement error; reliability; construct-level relationships; disattenuation; structural equation modeling*

Measurement error has been the focus of attention of both physical and social scientists for decades (Fuller, 1987; Lord & Novick, 1968). The potential deleterious effect of measurement error on substantive conclusions in scientific research has been well documented (e.g., Cronbach, 1947; Ree & Carretta, 2006; Schmidt & Hunter, 1996, 1999; Stanley, 1971; Thorndike, 1951). However, in the social sciences, many substantive researchers still tend to overlook or misunderstand the effects of measurement error on research results. Such practice is unfortunate, especially given that in most substantive areas, research has progressed beyond a focus on relationships between measures to a focus on relationships between constructs (J. P. Campbell, 1982). Yet, in addition to measurement error, there are other factors that contribute to the observed variances of measures of organizational and psychological constructs. As discussed later in this article,

Authors' Note: The authors thank Remus Ilies for allowing them to use data from his published study in the article. Please address correspondence to Huy Le, Department of Psychology, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816-1390; e-mail: hale@mail.ucf.edu.

these factors are specific to the scales but not theoretically relevant to the constructs of researchers' interests. The estimation of construct-level relationships is possible only when one controls for the distorting effects of these measurement artifacts.

Scientific theory includes explanations of the relationships among scientific constructs. These relationships are typically estimated as correlations, and observed correlations are generally downwardly biased as estimates of the construct-level correlations of interest because of measurement artifacts, which include measurement errors and other construct-unrelated sources of variances in measures. If properly applied, structural equation modeling (SEM) and confirmatory factor analysis (CFA) can correct for all sources of measurement artifacts; in practice, however, as described later, most applications of SEM/CFA fail to correct for at least two sources of measurement artifacts. Moreover, most studies do not employ SEM or CFA. In these studies, the corrections made are typically incomplete, causing these estimates of construct-level correlations to differ from those obtained when SEM or CFA are appropriately applied.

In this article, we propose a solution to this problem. Specifically, we address the issue in the following way: First, we briefly review the major sources of measurement artifacts under the conceptualization of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). We then present a new coefficient, the generalized coefficient of equivalence and stability (GCES), which calibrates all the major sources of measurement artifacts in self-report measures of organizational and psychological constructs and therefore can be used to estimate the true relationships between the constructs of interest. We next consider a simple, specific SEM model derived from the multitrait-multimethod (MTMM) literature that is conceptually equivalent to the GCES approach. Both correction procedures based on GCES and SEM make use of a special study design that enables controlling for the effects of all sources of measurement artifacts. We examine the performance of these two correction procedures across different conditions using the Monte Carlo simulation method. Because the procedures rely on basic assumptions of classical test theory, which may not hold in certain situations,[1] we further simulate data to examine the robustness of these methods. Finally, we suggest an approach to meta-analytically estimate the GCES for measures of well-established constructs. These estimates can be conveniently used by substantive researchers to correct for the effect of measurement artifacts in their studies in research situations where the assumption of uncorrelated measurement artifacts (transient error) is met. As discussed later, this represents an advantage of the GCES procedure as compared with the SEM procedure. As an illustration, we present an application of the approach for the case of overall job satisfaction, an important organizational construct.

The Multifaceted Nature of Measurement Artifacts in Self-Report Measures

Under classical test theory, an observed score (of a measurement object) is the sum of a true score and measurement errors. These sources of variance are assumed to be independent of each other, so the observed score variance of a measure across measurement objects is the sum of true score variance and the variances of measurement errors. For self-report measures used in many studies of psychological and organizational constructs,

there are three major sources of measurement error that contribute to the variance of observed scores: random response error, transient error, and item-specific factor error (Schmidt & Hunter, 1999; Schmidt, Le, & Ilies, 2003). The true score variance of a measure, however, does not purely reflect variance of the construct of interest (Schmidt, Viswesvaran, & Ones, 2000). Included in the true score variance of a measure are variances due to the construct of interest and other factors that are specific to the measure and not theoretically relevant to the construct. The effect of the latter sources of variation on construct-level relationship estimates is conceptually and functionally similar to those of measurement errors, that is, it attenuates the observed correlations between measures. As such, like measurement errors, they are measurement artifacts and are referred to as scale-specific factor error (Schmidt & Hunter, 1999; Schmidt et al., 2003).

Under generalizability theory (Cronbach et al., 1972), these sources of measurement artifacts can be conceptualized as different measurement facets. Specifically, here we have a three-facet, partially nested model with persons (p) being the object of measurement, and occasions (o), items (i), and scales (s) being the measurement facets. As mentioned above, items are nested within scales (measures), so its effect can be presented as $i : s$. This is the $p \times o \times (i : s)$ model in the terminology of generalizability theory. Accordingly, observed score variance across persons includes five major variance components corresponding to the sources of measurement in question, as shown in the following equation (Cronbach et al., 1972; Shavelson & Webb, 1991):[2]

$$Var(X) = Var(p) + Var(po) + Var(pi : s) + Var(ps) + Var(pos, poi : s, e). \quad (1)$$

where $Var(X)$ is observed score variance; $Var(p)$ is variance due to persons' differences in the construct of interest (construct variance); $Var(po)$ is variance due to interaction between persons and occasions (transient error variance); $Var(pi{:}s)$ is variance due to interaction between persons and scale items (variance due to item-specific factor error); $Var(ps)$ is variance due to interaction between persons and scales (variance due to scale-specific factor error); $Var(pos, poi{:}s, e)$ is the combination of variance due to random response error (e) and those due to three-way interactions between persons and other measurement facets.

The substantive nature of the three sources of measurement error (random response error, transient error, and item-specific factor error) has been discussed elsewhere (Cronbach, 1947; Schmidt et al., 2003). The existence of random response error and item-specific factor error in measures of psychological and organizational constructs is, in fact, well-accepted, as reflected by the universal use of coefficient alpha (coefficient of equivalence, CE) as *the* reliability coefficient of choice in the literature. In contrast, little is known about the two other sources of measurement artifacts in Equation 1. In the following section, we elaborate on the substantive meanings of transient error and scale-specific factor error and discuss why they may contribute to the observed variances of measures of psychological and organizational constructs.

Transient Error

Transient error results from moods, feelings, or mental states that are specific to a particular occasion (Becker, 2000; Cronbach, 1947; DeShon, 1998; Schmidt & Hunter, 1999;

Schmidt et al., 2003; Thorndike, 1949, 1951). According to generalizability theory, it is a source of measurement error resulting from the sampling of occasions. On a particular occasion, transient error exerts its effect on a participant's responses to all questionnaire items. However, this effect will be different on other occasions, because the mental states vary from occasion to occasion.[3] As such, the participant's score on the same measure may vary depending on his or her mental state on each measurement occasion. The variance in observed scores caused by variation in mood or mental states is transient error variance. This variation is considered measurement error variance if the construct in question is considered in the relevant theory to be temporally stable over specified time periods (e.g., ability, personality, general attitudes). If the construct is not considered stable (e.g., mood), then the variation is a part of the construct variance and therefore the concept of transient error does not apply.

Transient error almost certainly exists in measures of many organizational constructs such as job satisfaction and organizational commitment. Consider the case of job satisfaction. A person's response to a job satisfaction measure at any point in time is likely to reflect a combination of (a) the person's evaluation of the job environment and (b) daily fluctuation in his or her mood or mental state. Only the former reflects the construct of job satisfaction. The latter is not theoretically relevant to the construct of interest and should be treated as a source of measurement error. Some may argue that the daily mood fluctuation should be considered part of the job satisfaction construct. However, if this were the case, job satisfaction should always be defined at a specific point in time such that any interpretation of the construct should reference that time frame (e.g. "correlation between job performance and job satisfaction as measured on March 29, 2001"). Such a level of specificity lacks scientific generality. Schmidt and Hunter (1996; Scenario 19) state the following in the case of transient error in measures of organizational commitment:

> Suppose that organizational commitment feelings on a given day depend on the person's emotional state at that moment. In particular, suppose that commitment depends on the person's level of physical energy. Suppose that when people feel energetic, they look forward to work and have higher commitment feelings and that when they are sick and feel lethargic, they have lower commitment feelings toward work. Would any theorist really want to define the important construct of organizational commitment in such a way that commitment is lower if the person has a cold than if the person is well? Probably not. That is, it would be obvious that transient variation is measurement error and not part of the construct itself. (p. 216)

Scale-Specific Factor Error

In many research areas, several scales have been developed to measure the same theoretical construct (e.g., self-esteem, emotional stability, job satisfaction). These scales may or may not be based on the same theory or model—for example, Job In General (JIG; Ironson, Smith, Brannick, & Gibson, 1989) and Minnesota Satisfaction Questionnaire (MSQ full scale; Weiss, Dawis, England, & Lofquist, 1967) are two measures that are ostensibly based on different conceptualizations of the construct of overall job satisfaction—but they are nevertheless used by researchers to empirically examine the relationships between that construct and other constructs in its nomological network (e.g., job satisfaction as a determinant

of turnover; Hom, Caranikas-Walker, Prussia, & Griffeth, 1992). Results are typically not interpreted as being specific to the measures used in those studies (e.g., overall job satisfaction as measured by the JIG), but rather, conclusions are drawn about the construct in general. As such, the conclusions are assumed to be generalizable to all measures for the construct. Agreement in research findings obtained from different measures is therefore desirable, and theories empirically verified thereby would be considered as having strong (external) validity (T. D. Cook & Campbell, 1979). Such convergence underlies the critical multiplism approach to scientific inquiry (T. D. Cook, 1985; Houts, Cook, & Shadish, 1986). Arguably, scientific progress that furthers our understanding about the nomological networks of psychological and organizational constructs resides in findings being shared among different measures of the same construct, whereas those findings unique to a specific measure are probably of less scientific value. But what accounts for the commonality across different measures of the same construct? Or, putting it another way, which factor(s) results in findings specific to a certain measure that are not replicable by other measures of the same construct?

Equation 1 suggests an answer to the questions: True score variance of a measure, which is the difference between observed variance [$Var(X)$] and measurement error variances under classical test theory (i.e., random response error, transient error, and item-specific factor error variances), includes variance due to the construct [$Var(p)$] and that due to scale-specific factor error [$Var(ps)$]. The former component, $Var(p)$, explains findings about the constructs that are generalized across measures. The latter, $Var(ps)$, being specific to a measure, accounts for nongeneralizable findings and functions as other classical measurement errors in attenuating the observed correlations between the measure and those of other constructs.

An analogy can be drawn between item-specific factor error and scale-specific factor error. Just as every item within a scale has a specific factor that is considered measurement error in classical test theory, so every scale contains specific factor(s) that can be considered measurement error (Schmidt & Hunter, 1999). In effect, when generalizability theory is applied in this context (i.e., when scales are considered levels of a measurement facet under the generalizability theory model as suggested earlier), each scale is treated as a representative of the population of scales measuring the construct in question, just as items are considered as being drawn from the population of items under classical test theory (Nunnally & Bernstein, 1994). The idiosyncrasy (specific factor) of each scale creates a source of variance in observed scale scores that is unrelated to the underlying construct that the scales were constructed to measure, just as does the specific factor of an item discussed earlier. The resulting variance is scale-specific factor measurement error variance, which is variance due to interaction between persons and scales [$Var(ps)$ in Equation 1]. Hence, the research program associated with any one of these scales is ''trapped'' within the limitations of that scale and results are not readily generalizable.

For example, the literature contains numerous measures of verbal ability that were not constructed to be parallel forms of each other as defined in classical test theory (Lord & Novick, 1968). These measures were constructed by different researchers at different times, with no explicit attempt at making them parallel in the classical test theory sense. However, all researchers were working from the same general concept (theoretical model) of verbal ability. As long as such measures tap a single factor, for theory development

purposes, it is desirable to define the theoretical construct of verbal ability as the construct shared by all such measures. Suppose we use five such scales simultaneously to measure verbal ability, with the total or average score across these scales being the final observed verbal ability score. Such a multiple-measurement scale can be expected to have construct validity superior to that of single scale, but what is the appropriate reliability for this scale? One way to estimate reliability of the scale is by treating each scale as an item in a five-item test, then computing its CE (Schmidt et al., 2003). This reliability estimate assigns scale-specific factors to measurement error. Research findings obtained using this reliability estimate to correct for bias introduced by measurement error generalize to the population of all such measures of verbal ability. The theory constructed in this manner will, hence, have more scientific generality and, in that important sense, will be a better theory.

Another empirical example dramatically illustrating the importance of controlling this source of measurement artifacts is presented in Green, Goldman, and Salovey (1993). This study showed that the correlation between the constructs of negative effect and positive effect, when each was measured by a single measure, was only -0.34. This value was obtained by applying the SEM model with a single measure of each of the two constructs and with clusters of items from each individual measure being indicators; as discussed later, this is tantamount to correcting the observed correlation for unreliability using the classical CE (coefficient alpha). However, when multiple measures (different scales) were used as indicators for each construct in the SEM model (thereby accounting for factors specific to each scale, that is, scale-specific factor error), the correlation between these constructs was estimated as -0.85. A difference this large (-0.34 vs. -0.85) has major implications for theory development. The Green et al. (1993) study illustrates that scale-specific factor error can seriously distort substantive research conclusions if it is not appropriately accounted for. Conceivably, similar results may be found in other research situations. Yet, this source of measurement artifacts is generally unknown by many substantive researchers.

The concept of scale-specific factor error, although not explicitly discussed, is neither new nor novel in organizational research. There have been studies that employed different measures to operationalize an organizational construct, thereby controlling for the effects of scale-specific factor errors (e.g., Hom & Griffeth, 1991: job satisfaction; Harris & Bladen, 1994: role conflict, role ambiguity, role overload). The existence of scale-specific factor errors and its implications in estimating construct-level relationships are in fact implicit in the MTMM approach well-accepted by organizational researchers (cf. Doty & Glick, 1998).

Correcting for the Biasing Effects of Measurement Artifacts

The Disattenuation Formula

The concept of reliability and the associated rationale for correction for attenuation due to unreliability of measuring instruments is presented at length in psychometric texts (e.g., Cronbach et al., 1972; Lord & Novick, 1968; Nunnally & Bernstein, 1994). The fundamental concept is that measurement error biases observed correlations downward as estimates of the correlations between true scores of the constructs of interest and that it is possible to remove this bias by the use of appropriate correction. In classical measurement

theory, the formula for estimating the true-score correlation underlying the observed correlation between two measures (x and y) is

$$\rho_{xtyt} = \rho_{xy} / (\rho_{xx}\rho_{yy})^{1/2} \quad (2)$$

where ρ_{xy} is the observed correlation, ρ_{xtyt} is the correlation between the true scores for measures x and y, and ρ_{xx} and ρ_{yy} are the reliabilities of x and y, respectively.

The use of the appropriate reliability coefficient in Equation 2 is obviously critical to accuracy. If the reliability coefficient fails to account for all sources of measurement error, its use in the disattenuation formula will generally lead to undercorrecting for the downward biases that measurement error exerts on the observed correlation, underestimating the magnitude of relationship between the true scores underlying the measures. For estimating the true-score correlations from observed correlations of self-report measures, Schmidt et al. (2003) recommended the coefficient of equivalence and stability (CES), which controls for the three sources of measurement errors (random response error, transient error, and item-specific factor error). True score correlations, however, are biased estimates of the construct-level relationships because they are still "contaminated" by scale-specific factor error, whose variance contributes to true score variance (see Equation 1). Thus, to estimate the construct-level correlation, a new coefficient is needed to further remove the biasing effect due to scale-specific factor error.

The Generalized Coefficient of Equivalence and Stability

In classical test theory, reliability can be estimated by taking the correlation between parallel scales when its basic assumptions (i.e., true score and measurement error are uncorrelated, measurement errors are uncorrelated, and there are parallel forms having the same true score variance and error variance) are met. Schmidt et al. (2003) showed that the CES obtained by correlating parallel forms administered on different occasions accounts for all major sources of measurement error in self-report measures as defined by classical test theory and, therefore, is the most appropriate type of reliability to be used in the disattenuation formula for estimating the true-score correlations between measures.

A similar procedure can be suggested to estimate construct-level relationships based on the disattenuation formula. From Equation 1, it can be shown (see Appendix A for details) that the correlation between two different measures of the same construct, when they are administered on two different occasions, provides an estimate of the proportion of construct variance [$Var(p)$] to the observed variance [$Var(X)$]. We call that proportion GCES, as it accounts for all classical sources of measurement error (as does the CES) and scale-specific factor error, thereby ensuring the generalizability of findings across measures of the same constructs. The GCES can then be used in Equation 2 to correct for the effects of all major sources of measurement artifacts in self-report measures to estimate construct-level relationships.

Assumptions Underlying the GCES Correction Procedure

The basic assumption of the GCES correction procedure is implied in Equation 1, which stipulates that observed scores are the sum of different, independent measurement

facets (cf. the additive model; Bagozzi & Yi, 1990; Becker & Cote, 1994). This assumption is central in classical test theory and generalizability theory. The independence between construct and measurement artifacts is, in fact, based on the conceptualizations of these measurement facets. For this model, the independence assumption is inherent in the definitions of the measurement artifacts. In particular, random response error is defined as a source of variation specific to a moment, unrelated to the construct being measured (Schmidt et al., 2003). By the same token, transient error, item-specific factor error, and scale-specific factor error are defined as sources of variations specific to an occasion, item, and scale, respectively. There have been suggestions that these sources of measurement artifacts could be spuriously related to the construct (cf. James, Demaree, Mulaik, & Ladd, 1992). Such construct-artifact relationships form the basis of the multiplicative model, which is an alternative model for the traditional additive effect model (Bagozzi & Yi, 1990). Empirical evidence, however, shows that the additive model is generally appropriate for organizational constructs (Becker & Cote, 1994).

In addition to the additive effect model assumption, there are three important assumptions needed for the GCES correction procedure. The first assumption is about the nature of the differences between measures of a construct. As discussed earlier, the measures are assumed to reflect the same theoretical construct, so their observed differences are merely due to the idiosyncratic characteristics of the measures. The second assumption requires that there exist "parallel measures" (that is, measures having the same proportion of $Var(p)$ to $Var(X)$; see Appendix A for details) for a construct. This is analogous to the assumption of parallel scales in classical test theory, although the meaning of parallelism is different in this context. The third assumption stipulates that the measurement artifacts between two measures are uncorrelated. We discuss these assumptions in detail next.

Nature of observed differences between measures. This assumption is crucial for our conceptualization of scale-specific factor error as a measurement artifact. As discussed in the "Scale-Specific Factor Error" section, the same theoretical construct is assumed to underlie different measures, hence, any differences observed between them should arise from the idiosyncrasies of the measures, not from any substantive reasons. Arguably, this assumption is satisfied for measures of most well-established psychological and organizational constructs (e.g., general mental ability, the Big Five personality traits, overall job satisfaction measures). Some constructs, however, are conceptualized differently by different researchers, who developed measures for the constructs based on their own conceptualizations. Constructs underlying these measures, although referred to by the same name, are actually not the same. For example, organizational commitment was conceptualized as either a unidimensional construct (Mowday, Steers, & Porter, 1979) or a multidimensional construct (Meyer, Allen, & Smith, 1993). Correlations between measures for these constructs are attenuated by both their substantive differences and scale-specific factor error. It is therefore not possible to estimate the GCES of these measures from their correlations, and hence, correction based on the GCES procedure cannot be implemented.

Existence of parallel measures. As shown in Appendix A, the correlation between two scales measuring the same construct on different occasions is the geometric mean of the GCESs of the scales. If we assume that scales developed to measure the same construct

have similar values of GCES, that is, their proportions of construct variance [*Var*(*p*)] to observed variance [*Var*(*X*)] are the same, their correlation is equal to the GCES of the scales. Although this assumption may appear rather restrictive and therefore difficult to satisfy, we believe that for most well-established constructs, factor structures of their measures are reasonably similar. It is likely that such measures, although developed by different researchers at different times, have been based on the same theories of the underlying constructs, resulting in similar factor structures.

Nevertheless, given this seemingly restrictive assumption, it is important to assess the accuracy of the GCES correction method when the assumption is violated. Thus, we later carry out additional simulation studies examining the accuracy of the procedure when the variance components of measures of the same construct are not equal.

Uncorrelatedness among the measurement artifacts. This assumption is needed to allow the disattenuation formula (Equation 2) to work. It means that the observed correlation between two measures is solely due to their underlying constructs, not measurement artifacts. In other words, measurement artifacts only attenuate, not inflate, the observed relationships. Conceivably, this assumption holds for most measurement artifacts by virtue of their definitions. For example, random response error cannot be correlated with any other artifacts because it is specific to a moment. By the same token, item-specific factor error and scale-specific factor error of a measure, being specific to the measure in question and its items, should not be correlated to those of another measure. The only potential exception is transient error. It is possible that the transient factor (mood) of a person exerts a similar effect on his or her responses to measures of related constructs, say, overall job satisfaction and organizational commitment. In other words, transient errors of measures for different, but related, constructs can be correlated, leading to violation of the assumption of uncorrelated measurement artifacts. Because this violation is likely to occur in certain situations, the correction procedure based on GCES should take it into account. We discuss such an adapted procedure in the following section.

Estimating Construct-Level Relationships Using the GCES When Transient Errors Are Correlated

Although transient errors of measures for related constructs may be correlated when the measures are administered on the same occasion, they are certainly not correlated when the measures are administered on different occasions (assuming that the two occasions are separated by an appropriate interval). Thus, to avoid the confounding effect of correlated transient error, we need to use the correlations between measures administered on two different occasions in the disattenuation formula (Equation 2). We illustrate the correction procedure below using a mock data set.

Let's consider a situation where we are interested in estimating correlation between two constructs, *K* and *L*. Assume that we have two measures for each construct, *A* and *B* for construct *K*, and *C* and *D* for construct *L*. Measures *A* (for construct *K*) and *C* (for construct *L*) are administered on occasion 1 to a sample of 500 participants. On occasion 2, measures *B* and *D* are administered to the same sample. We simulated data following the model shown in Equation 1 with values of the parameters presented in Table 1a. Table 1b

Table 1a
Parameters Used to Generate the Demonstration Data Set

	ρ_{KL}	$\rho_{p_{K^o}, p_{L^o}}$	*Var*(*p*)	*Var*(*po*)	*Var*(*pi:s*)	*Var*(*ps*)	*Var*(*e*)
Scales *A*, *B* (construct *K*)	.50	.60	.60	.10	.10	.10	.10
Scales *C*, *D* (construct *L*)			.50	.05	.15	.20	.10

Note: ρ_{KL} = true correlation between constructs *K* and *L*; $\rho_{p_{K^o}, p_{L^o}}$ = correlation between transient errors in measures of constructs *K* and *L*; *Var*(*p*) = construct variance; *Var*(*po*) = transient error variance; *Var*(*pi:s*) = item-specific factor error variance; *Var*(*ps*) = scale-specific factor error variance; *Var*(*e*) = random response error variance (refer to Equation 1 in the text).

shows the correlation matrix for the measures obtained from the sample. The correlation between *A* (time 1) and *B* (time 2) ($r_{A1B2} = .557$, Table 1b) provides an estimate for the GCES for measures of construct *K*. Similarly, the correlation between *C* and *D* ($r_{C1D2} = .424$) estimates the GCES for measures of construct *L*. Correlations between measures of the two constructs administered at the same time ($r_{A1C1} = .262$ and $r_{B2D2} = .299$) are simultaneously attenuated by the measurement artifacts and inflated by the correlations between transient errors ($\rho_{p_{K^o}, p_{L^o}} = .600$), so they cannot be used in the disattenuation formula to estimate the correlation between constructs *K* and *L*.[4] The correlation between *A* at time 1 and *D* at time 2 ($r_{A1D2} = .237$), and that between *B* at time 2 and *C* at time 1 ($r_{B2C1} = .246$) are not inflated by correlated transient error. The average of these two correlations (.242) can be used in the disattenuation formula together with the estimates for GCES of the measures (.557 and .424 for measures of constructs *K* and *L*, respectively) to estimate the construct-level relationship between *K* and *L*. The resulting value, .498, virtually matches the parameter used to simulate the data ($\rho_{KL} = .500$, Table 1a).

SEM and the Estimation of Construct-Level Relationships

Conceptual equivalence between the SEM and GCES approaches. SEM (and CFA) has been suggested as a tool to correct for the effects of measurement errors and construct-unrelated factors to estimate the relationships between constructs (Cohen, Cohen, Teresi, Marchi, & Velez, 1990; Doty & Glick, 1998; Marsh & Hocevar, 1988; Williams, 1995). It can be seen that the process of using GCES in the disattenuation formula to account for multiple sources of measurement artifacts described in the previous section is conceptually similar to that employed in SEM. In SEM (and CFA), when a construct (latent variable) is operationalized by several indicators, it is defined to be what is commonly shared by the indicators. If the indicators include different measures of the construct administered on different occasions, effects of all the four major sources of measurement artifacts discussed in this article (random response error, transient error, item-specific factor error, and scale-specific factor error) will be accounted for (i.e., will be delegated to residuals and therefore excluded from the construct), because they are unique to each indicator. As a consequence, correlations among constructs (latent variables) estimated in SEM are conceptually similar to those obtained when GCES is used in the disattenuation formula to correct for the effects of measurement artifacts.

Table 1b
Correlation Matrix of Measures of Constructs *K* and *L*
From the Demonstration Dataset

	A1	*B2*	*C1*	*D2*
A1	1.000			
B2	.557	1.000		
C1	.262	.246	1.000	
D2	.237	.299	.424	1.000

Note: $N = 500$. Data were generated based on the parameters shown in Table 1a. *A1* = scale *A* of construct *K* administered on occasion 1; *B2* = scale *B* of construct *K* administered on occasion 2; *C1* = scale *C* of construct *L* administered on occasion 1; *D2* = scale *D* of construct *L* administered on occasion 2.

The sampling design of indicators in SEM determines which sources of measurement artifacts the SEM model can control for. When items of a measure administered on a single occasion are split into clusters to form indicators (e.g., as in Brooke, Russell, & Price, 1988; Mathieu & Farr, 1991), the SEM model can account only for the effects of item-specific factor error and random response error, as with the use of CE in the disattenuation formula. Alternatively, if the indicators include several different scales measuring the same constructs administered on the same occasion, the construct-level correlation estimated by SEM will control for the attenuating effects of random response error, item-specific factor error, and scale-specific factor error, but not transient error (e.g., Cohen et al., 1990).

There are two SEM models that are especially relevant to this topic. They are the latent state-trait model (Schmitt & Steyer, 1993) and the well-known MTMM model (Campbell & Fiske, 1959). We discuss the models next to determine the appropriate SEM procedure that is equivalent to the GCES procedure discussed in the previous section in terms of controlling for measurement artifacts to estimate construct-level correlations.

The latent state-trait model. Schmitt and Steyer (1993) suggested the latent state-trait model, which decomposes the observed variance of measures into variances due to trait (construct), state (transient error), method-specific factors (scale-specific factor error or item-specific factor error), and residuals (random response error). This model highlights the importance of separating stable constructs from temporal state (transient error) and provides a means to achieve it. The researchers, however, did not explicitly discuss the meaning of the method-specific factors. As discussed earlier, these factors can either be item-specific factor error or scale-specific factor error, depending on which indicators are used in the model. Marsh and Grayson (1994) also presented several alternative SEM models and demonstrated how they can be used to partition observed variance into variances due to the common factor, time-specific factor, item-specific factor, and residuals. It could be seen that the latent state-trait model suggested by Schmitt and Steyer is similar to one of the models presented by Marsh and Grayson. Although Marsh and Grayson also did not mention scale-specific factor error in their study, they noted that the models could be expanded to cover other factors. None of these studies, however, explicitly discusses how the models could be used to estimate the construct-level relationships by controlling for the effect of measurement artifacts in observed correlations.

The MTMM approach. The MTMM approach (Campbell & Fiske, 1959) allows examination of construct validity of measures by accounting for and comparing different sources of variances (traits and measurement methods) contributing to the measures' observed variances. MTMM, as the name indicates, requires concurrently examining a number of different traits and different measurement methods. Convergence across measures applying different measurement methods on a trait indicates the construct validity of the measures. As such, MTMM is directly relevant to our conceptualization of GCES and measurement artifacts.

In MTMM CFA models, variances in observed measures are decomposed into methods and traits, allowing for estimation of correlations among traits (cf. Doty & Glick, 1998). Method factors are those shared by similar methods applied by measures for different constructs. Research on MTMM (e.g., Doty & Glick, 1998; Williams, Cote, & Buckley, 1989) consistently confirms the existence of method factors in measures of psychological and organizational constructs. This would mean that correlations between measures of different constructs based on the same measurement method may be inflated because of the shared method (Doty & Glick, 1998). As the result, the assumption of uncorrelated measurement artifacts suggested for our GCES model appears not tenable for measures based on the same measurement method. Doty and Glick examined the effects of correlated method factors in MTMM studies and concluded that (a) the methods are correlated and (b) such correlation leads to overestimation of the construct-level correlations. These findings appear to seriously question the assumptions of uncorrelated measurement artifacts underlying the GCES estimation procedure discussed earlier and even the viability of the procedure.

To better understand the implications of such findings for the GCES assumption, it is important to examine the correspondence between common method factors in MTMM and measurement artifacts in this GCES model. Method factors in MTMM pertain to measurement techniques that create variances in measures that are not related to the construct of interest (Doty & Glick, 1998; Kenny, 1995). As such, the common method factors in MTMM appear to be similar to scale-specific factor error discussed here thus far. However, although related, the two concepts are not totally overlapping. There are three basic types of method factors in MTMM studies (Conway, 1998; Doty & Glick, 1998; Kenny, 1995): (a) rater as method (different raters provide assessments for the construct in question), (b) instrument-based methods (different measures are used by the same raters), and (c) temporal methods (same measures are used on multiple occasions). The first type, rater as method, is not applicable here, as we are concerned about self-report measures. The second type (instrument-based methods) is relevant to our conceptualization of item- and scale-specific factors. The third type of method factors, temporal methods, directly corresponds to our conceptualization of transient error. A close examination of MTMM studies examined in Doty and Glick (1998) suggests a reason that the method factors as operationalized in those studies are likely to be correlated: Most studies there were either administered at the same time or the difference in instrument-based methods examined in those studies is minimal (measures using question items with the same contents, with only different response scales). As discussed earlier, transient error for measures of related constructs administered on the same occasions can be correlated. And measures with similar question items are likely to be similarly influenced by item-specific factor errors, leading

to intermethod correlation. In fact, there is only one study rated by Doty and Glick (Jaffe & Nebenzahi, 1984, cited in Doty & Glick, 1998) as reasonably different in terms of both temporal and instrument-based methods. The authors found that method factors are uncorrelated in that study (.007). Taken together, Doty and Glick's findings confirm existence of measurement artifacts and further suggest the problem of correlated transient errors, which need to be taken into account to estimate correlations between constructs.

As such, it can be seen that MTMM research dovetails nicely with the conceptualization of measurement artifacts presented in this article. MTMM methods, when the method factors are appropriately operationalized to account for the effects of measurement artifacts, can be used to estimate relationships between constructs (cf. Doty & Glick, 1998). In the following section, we consider a specific MTMM method, the uniqueness model (CU; Conway, 1998; Lance, Noble, & Scullen, 2002; Marsh, 1989), as a procedure that is equivalent to the GCES procedure discussed earlier for estimating construct-level correlations.

A simple SEM model to estimate construct-level relations. As mentioned earlier, the sampling of indicators for constructs in SEM models has important implications in the estimation of construct-level relationships. Specifically, for a construct, it is essential that its indicators include several different measures and occasions. For simplicity and practicality, measures can be nested within occasions, that is, different measures (for a construct) are administered across occasions.[5] The simplest example for this design is the situation examined in the earlier section where we estimated the relation between two constructs (K and L), each having two measures (A, B for construct K, and C, D for L) administered on two occasions (A and C at occasion 1, and B and D at occasion 2; see Tables 1a and 1b). The design makes efficient use of researchers' resources as it requires only administering the smallest number of measures (one for each construct) each time.

Figure 1 illustrates the model underlying data in Table 1b. A response of participant i ($i = 1$ to N) to a measure of construct j($j = K$ or L) on occasion k($k = 1, 2$) can be decomposed into two components representing effects due to (a) the underlying construct and (b) measurement artifacts (transient error, scale-specific factor error, item-specific factor error, and random response error combined):

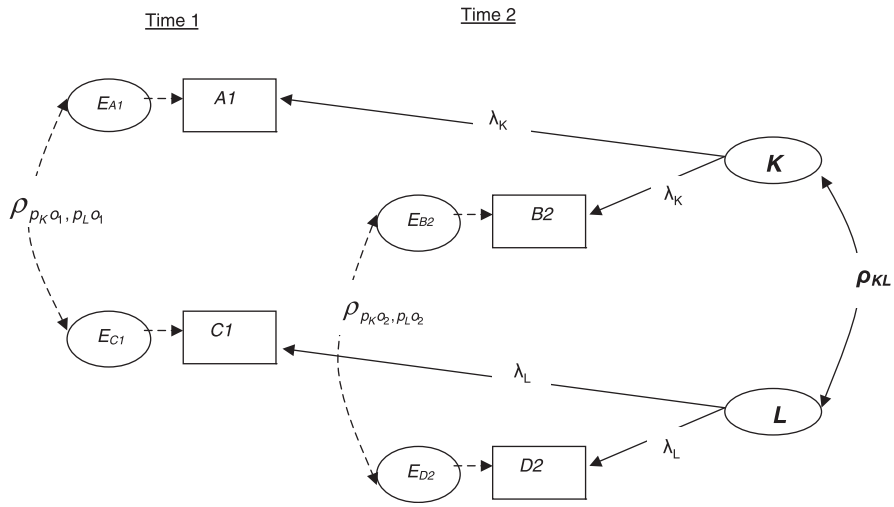
$$y_{ijk} = \lambda_{jk}\eta_{ijk} + \varepsilon_{ijk}, \tag{3}$$

where y_{ijk} = response of person i on a measure of construct j on occasion k; λ_{jk} = the path from construct j to its measure administered on occasion k (e.g., for $j = K$ and $k = 1$, it is the path from construct K to scale A); η_{ijk} = value of person i on construct j on occasion k; ε_{ijk} = the component in person i's response due to measurement artifacts (including transient error, scale-specific factor error, item-specific factor error, and random response error). The variance-covariance matrix for the model in Equation 3 is

$$\Sigma = \Lambda\Phi\Lambda' + E, \tag{4}$$

where Σ = the 4×4 (jk by jk) variance–covariance matrix among the indicators of the model (assuming the scores are standardized, we have the corresponding correlation matrix shown in Table 1b); Λ = the 4×2 (jk by j) matrix reflecting the loadings of

Figure 1
The Structural Equation Model for the Simulated Data



construct by occasion (confounded with measures); Φ = the 2 × 2 (j by j) matrix reflecting the variances/covariances between K and L (for standardized solution, this is the construct-level correlation matrix of interest); E = the 4 × 4 (jk by jk) matrix reflecting the variances/covariances among the residuals of the indicators (measurement artifacts).

The residual variance-covariance matrix E is the symmetric matrix with values in the diagonal being equal to the total variances of all measurement artifacts for the four scales ($A1, B2, C1$, and $D2$, respectively). For the off-diagonal cells, $e_{jk,\,j'k'} = 0$ (i.e., fixed at zero) when $k \neq k'$, and $e_{jk,\,j'k'} \neq 0$ (i.e., freely estimated) when $k = k'$(and $j \neq j'$). The nonzero values in the off-diagonal cells of matrix E reflect correlated transient errors. Figure 1 illustrates the model. Note that for the current model, factor (construct) loadings for measures of the same constructs are constrained to be equal (that is, $\lambda_{K1} = \lambda_{K2} = \lambda_K$, and $\lambda_{L1} = \lambda_{L2} = \lambda_L$). These constraints are needed for the model to be determinable. They are also consistent with the assumptions discussed earlier.

For the demonstration data set shown in Tables 1a and 1b, we use Proc CALIS (SAS 9.1) to estimate the construct-level relationship between K and L based on the model in Figure 1. The model shows very good fit (root mean square error of approximation [RMSEA] = .001, chi-square = .02, p = .88 with df = 1, comparative fit index [CFI] = .999). Estimated correlation between the two constructs ($\hat{\rho}_{KL} = .496$) is very similar to that obtained from the GCES procedure reported in the previous section, which is very close to the parameter used to simulated the data ($\rho_{KL} = .500$). These results appear to show that the two procedures (GCES and SEM) introduced here perform equally well in estimating construct-level correlation by accounting for the effects of measurement artifacts. However, because these results are based on one single data set with a specific factor structure (i.e., the proportions of variances of measurement facets to observed variance), it is not clear if the findings are

generalizable to other situations. In the following section, we attempt to answer this question by examining performance of the procedures more thoroughly using the Monte Carlo data simulation technique. Furthermore, as discussed earlier, the procedures are based on assumptions that might be violated in certain situations, so we need to examine their robustness when these assumptions are violated.

The Monte Carlo Simulation

Conditions examined. We simulated data and carried out analyses separately for each procedure as described in the previous sections. There is a total of nine conditions examined. Design of these conditions is similar to the example examined in the previous sections. That is, it is assumed that there are two constructs; each has two scales administered on two different occasions. All of the scales are administered on two occasions to the same 500 participants. Thus, for each data set, there are four data points for each of the participants (2 constructs × 2 scales confounded with 2 occasions).

The population correlation between the two constructs is set to be .50 for all conditions. Details of other parameters used to simulate data are shown in Table 2. The first three conditions (conditions 1, 2, and 3) are similar to the example presented in the previous sections in terms of factor structures (proportions of construct variance and measurement artifact variances to observed variance). In these conditions, the factor structures for two scales measuring the same construct are similar, which means that they meet the assumption of parallelism discussed earlier (the GCES for the scales are the same). Correlation between transient errors of measures for different constructs when they are administered on the same occasions ($\rho_{p_{Ko}, p_{Lo}}$) is equal to 0 in condition 1, so this condition further meets the assumption of uncorrelated measurement artifacts. The correlation is moderate in condition 2 (.30) and relatively large in condition 3 (.60). Examining results in conditions 2 and 3 allows us to evaluate how well the procedures perform (that is, how successful our adaptations introduced into the procedures work) when the assumption of uncorrelated measurement artifacts is violated. The next three conditions (4, 5, and 6) are based on scales with different factor structures. In these conditions, scales for the same construct have different factor structures, leading to violation of the parallelism assumption. The ratio of GCES for two measures of the same construct can be used as an index for the extent to which the parallelism assumption is violated. Specifically, the GCES for measures of construct *K* are .60 and .50 (ratio 1.20) and those for construct *L* are .70 and .45 (ratio 1.6). Like the first conditions, the correlation between transient errors in conditions 4, 5, and 6 varies from .00 (perfectly meeting the assumption of uncorrelated measurement artifacts) to .60 (seriously violating the assumption). Finally, conditions 7, 8, and 9 include scales with different factor structures (ratio of GCES for measures of construct *K* is 1.5, and that for construct *L* is 2.0), providing another opportunity to examine the robustness of the two estimation procedures when the parallelism assumption is violated.

For each condition, 500 data sets were simulated. Results (correlations between constructs *K* and *L*) were estimated for each data set. We calculated the means and standard deviations (which are standard errors of the mean estimates) of these estimates across 500 data sets within each condition. The mean estimates were then compared with the true

Table 2
Simulation Parameters

Simulation Condition		$\rho_{p_{K^o}, p_{L^o}}$	Measures for Construct K					Measures for Construct L				
			Var(*p*)	*Var*(*po*)	*Var*(*pi:s*)	*Var*(*ps*)	*Var*(*e*)	*Var*(*p*)	*Var*(*po*)	*Var*(*pi:s*)	*Var*(*ps*)	*Var*(*e*)
Condition 1	Scale 1	.00	.60	.10	.10	.10	.10	.50	.05	.15	.20	.10
	Scale 2		.60	.10	.10	.10	.10	.50	.05	.15	.20	.10
Condition 2	Scale 1	.30	.60	.10	.10	.10	.10	.50	.05	.15	.20	.10
	Scale 2		.60	.10	.10	.10	.10	.50	.05	.15	.20	.10
Condition 3	Scale 1	.60	.60	.10	.10	.10	.10	.50	.05	.15	.20	.10
	Scale 2		.60	.10	.10	.10	.10	.50	.05	.15	.20	.10
Condition 4	Scale 1	.00	.60	.05	.10	.15	.10	.45	.10	.08	.22	.15
	Scale 2		.50	.15	.20	.10	.05	.70	.05	.05	.15	.05
Condition 5	Scale 1	.30	.60	.05	.10	.15	.10	.45	.10	.08	.22	.15
	Scale 2		.50	.15	.20	.10	.05	.70	.05	.05	.15	.05
Condition 6	Scale 1	.60	.60	.05	.10	.15	.10	.45	.10	.08	.22	.15
	Scale 2		.50	.15	.20	.10	.05	.70	.05	.05	.15	.05
Condition 7	Scale 1	.00	.60	.10	.10	.10	.10	.70	.15	.05	.05	.05
	Scale 2		.40	.10	.20	.20	.10	.35	.15	.20	.25	.05
Condition 8	Scale 1	.30	.60	.10	.10	.10	.10	.70	.15	.05	.05	.05
	Scale 2		.40	.10	.20	.20	.10	.35	.15	.20	.25	.05
Condition 9	Scale 1	.60	.60	.10	.10	.10	.10	.70	.15	.05	.05	.05
	Scale 2		.40	.10	.20	.20	.10	.35	.15	.20	.25	.05

Note: True correlation between constructs K and $L(\rho_{KL})$ is equal to .50 for all conditions. $\rho_{p_{K^o}, p_{L^o}}$ = correlation between transient errors of measures of the two constructs; *Var*(*p*) = construct variance (this is also the generalized coefficient of equivalence and stability because the observed variances are standardized); *Var*(*po*) = transient error variance; *Var*(*pi:s*) = item-specific factor error variance; *Var*(*ps*) = scale-specific factor error variance; *Var*(*e*) = random response error variance.

value of ρ_{KL} (.50) to examine potential biases. A series of SAS macros (SAS 9.1) was written to simulate data and estimate construct-level relationships.[6]

Results. All of the SEM analyses converged and had excellent fit (mean CFI > .99, mean RMSEA < .03, mean standardized root-mean-square residual [SRMR] < .015, mean goodness-of-fit index [GFI] > .99), despite the fact that the models were misspecified in conditions 4 through 9 (the paths from constructs to measures for the same constructs are constrained to be equal even when the parallelism assumption is violated). Table 3 presents the results. As can be seen therein, the two procedures equally provide very accurate estimates for the construct-level correlation. Violation of the uncorrelatedness assumption (in conditions 2, 3, 5, 6, 8, and 9) does not seem to affect performance for two procedures. When the assumption of parallelism is violated (conditions 4–9), both procedures slightly overestimate the construct-level correlations (.40% to 1.40% overestimation for the GCES procedure and .60% to 1.80% for the SEM procedure, compared with .20% to .60% overestimation for both procedures when the assumption is met in conditions 1-3). The biases are negligible in practice. The standard errors of estimates for the construct-level correlations are exactly the same for both procedures (averaged .057 across nine simulation conditions). Those results indicate that both the GCES and SEM procedures perform equally well in terms of accuracy and efficiency. Overall, our simulation results provide strong evidence that the two procedures introduced here are equivalent and can be used interchangeably to address the issue of estimating correlations between constructs in research.

Further examining the positive biases when the parallelism assumption is violated. The finding that both procedures consistently yield accurate estimates for the construct-level correlations is encouraging, but such a finding could have been due to the peculiarity of the values used in our simulations. To rule out this possibility and more thoroughly investigate the robustness of the GCES procedure, we analytically derive a formula to calculate the bias resulting from using the GCES procedure when its assumption of parallelism of measures for the same construct is violated. As shown in Appendix B, the construct-level correlation estimated by the GCES procedure is the function of (a) the true correlation between the constructs and (b) the GCESs of the scales involved:

$$\hat{\rho}_{KL} = \rho_{KL}\left(\frac{\sqrt{GCES_A}+\sqrt{GCES_B}}{2\sqrt[4]{GCES_A GCES_B}}\right)\left(\frac{\sqrt{GCES_C+GCES_D}}{2\sqrt[4]{GCES_C GCES_D}}\right). \quad (5)$$

where $\hat{\rho}_{KL}$ = construct-level correlation between K and L estimated by the GCES procedure;

ρ_{KL} = true correlation between K and L;
$GCES_A$, $GCES_B$ = the GCESs of scales A and B of construct K; and
$GCES_C$, $GCES_D$ = the GCESs of scales C and D of construct L.

We apply Equation 5 to different combinations of values of GCES for measures A and B of construct K and present the results in Table 4 (to simplify the presentation, we assume $GCES_C = GCES_D = 1.00$). In the table, we systematically vary the GCES value of scale A from .800 to .500 and that of scale B from .700 to .400. Three values of the correlation between K and L (.300, .500, and .700) are examined. We believe that the values

Table 3
Simulation Results

Condition	Assumption Violated	GCES Procedure			SEM Procedure		
		M	*SE*	% Bias	*M*	*SE*	% Bias
1	None	.503	.058	0.60	.503	.058	0.60
2	Uncorrelatedness	.501	.057	0.20	.501	.057	0.20
3	Uncorrelatedness	.501	.056	0.20	.503	.053	0.60
4	Parallelism	.505	.055	1.00	.508	.056	1.60
5	Parallelism + Uncorrelatedness	.506	.055	1.20	.509	.056	1.80
6	Parallelism + Uncorrelatedness	.502	.055	0.40	.505	.055	1.00
7	Parallelism	.502	.059	0.40	.503	.059	0.60
8	Parallelism + Uncorrelatedness	.502	.060	0.40	.503	.060	0.60
9	Parallelism + Uncorrelatedness	.507	.059	1.40	.508	.060	1.60
Average across conditions		.503	.057	0.64	.505	.057	0.96

Note: True correlation (ρ_{KL}) = .500 for all conditions. Uncorrelatedness = the assumption of uncorrelated measurement artifacts is violated; Parallelism = the assumption that the two scales for the same construct have the same GCES is violated; Parallelism + Uncorrelatedness = both assumptions are violated; GCES = generalized coefficient of equivalence and stability; SEM = structural equation modeling; *M* = mean of construct-level correlation estimates across 500 data sets; *SE* = standard error (standard deviation of the correlation estimates across 500 data sets); % Bias = percentage bias = 100*(M – .500)/.500.

included in Table 4 adequately cover most situations that researchers may encounter in practice. The second column from the right in Table 4 ($\hat{\rho}_{KL}$) shows the correlation between *K* and *L* estimated by the GCES procedure. The last column (% Bias) shows percentage bias when the assumption of equivalence of measures for the same construct is violated. As can be seen, despite the violation, the GCES procedure provides very accurate estimates of the correlation between *K* and *L*. The maximum bias, which is only 1.51%, occurs when $GCES_A = .800$ and $GCES_B = .400$. Arguably, situations where scales of the same construct are substantially different in the GCES are rare in practice. These results confirm the robustness of the GCES procedure: It can provide reasonably accurate estimates of the construct-level correlation despite serious violation of its major assumption about the equivalence of measures for the same construct.

Extending the GCES Approach by Means of Meta-Analysis

As demonstrated in the earlier section, the two procedures, GCES and SEM, are conceptually and computationally equivalent, so either can be used to account for the effects of measurement artifacts and estimate the relationships between constructs. The typical study, however, does not contain data on enough measures to allow implementation of either of these approaches. The normal estimation of the GCES and specification of the SEM model both require repeated administrations of different measures of the constructs of interest. Many studies, in fact, include only one measure of each construct. What is therefore needed is an approach that enables correction for all four sources of measurement

Table 4
Biases in the Estimated Correlations Using the Generalized Coefficient of Equivalence and Stability (GCES) Procedure When Its Assumption Is Violated

ρ_{KL}	$GCES_A$	$GCSE_B$	r_A	r_B	$\bar{r}$	$\overline{GCES}$	$\hat{\rho}_{KL}$	% Bias
.300	.800	.700	.268	.251	.260	.748	.300	0.06
.300	.800	.600	.268	.232	.250	.693	.301	0.26
.300	.800	.500	.268	.212	.240	.632	.302	0.69
.300	.800	.400	.268	.190	.229	.566	.305	1.51
.300	.700	.600	.251	.232	.242	.648	.300	0.07
.300	.700	.500	.251	.212	.232	.592	.301	0.35
.300	.700	.400	.251	.190	.220	.529	.303	0.98
.300	.600	.500	.232	.212	.222	.548	.300	0.10
.300	.600	.400	.232	.190	.211	.490	.302	0.51
.300	.500	.400	.212	.190	.201	.447	.300	0.16
.500	.800	.700	.447	.418	.433	.748	.500	0.06
.500	.800	.600	.447	.387	.417	.693	.501	0.26
.500	.800	.500	.447	.354	.400	.632	.503	0.69
.500	.800	.400	.447	.316	.382	.566	.508	1.51
.500	.700	.600	.418	.387	.403	.648	.500	0.07
.500	.700	.500	.418	.354	.386	.592	.502	0.35
.500	.700	.400	.418	.316	.367	.529	.505	0.98
.500	.600	.500	.387	.354	.370	.548	.501	0.10
.500	.600	.400	.387	.316	.352	.490	.503	0.51
.500	.500	.400	.354	.316	.335	.447	.501	0.16
.700	.800	.700	.626	.586	.606	.748	.700	0.06
.700	.800	.600	.626	.542	.584	.693	.702	0.26
.700	.800	.500	.626	.495	.561	.632	.705	0.69
.700	.800	.400	.626	.443	.534	.566	.711	1.51
.700	.700	.600	.586	.542	.564	.648	.701	0.07
.700	.700	.500	.586	.495	.540	.592	.702	0.35
.700	.700	.400	.586	.443	.514	.529	.707	0.98
.700	.600	.500	.542	.495	.519	.548	.701	0.10
.700	.600	.400	.542	.443	.492	.490	.704	0.51
.700	.500	.400	.495	.443	.469	.447	.701	0.16

Note: ρ_{KL} = true correlation between constructs K and L; $GCES_A$ = GCES of scale A of construct K; $GCES_B$ = GCES of scale B of construct K; r_A = observed correlation between scale A and the measure of construct L (GCES of the measure of construct L is assumed to be 1.00); r_B = observed correlation between scale B and the measure of construct L; $\bar{r}$ = average of the observed correlations between measures of construct K and L; $\overline{GCES}$ = correlation between scales A and B (this is the estimate of GCES of the scales); $\hat{\rho}_{KL}$ = estimate of the correlation between K and L by the GCES method; % Bias = percentage bias of the estimate.

artifacts in self-report measures in the absence of such a demanding study design. Such an approach would allow all researchers—not just those whose data allow the use of the SEM model and normal estimation of GCES as mentioned above—access to accurate estimates of construct-level correlations.

Below, we present an alternative approach for approximating the value of GCES by means of meta-analysis. Specifically, this approach uses cumulative knowledge in the literature about properties of measures of a construct to estimate the upper bound of GCES

for those measures through use of information derived from meta-analysis. It should be noted, however, that the approach described here requires the assumption of uncorrelated transient error (as noted earlier in the "Uncorrelatedness Among the Measurement Artifacts" section) because it is not possible to apply the adjustment mentioned in the "Estimating Construct-Level Relationships Using the GCES When Transient Errors Are Correlated" section without repeated measures design. Conceivably, this assumption can be met in most situations where the constructs under consideration are theoretically distinct (e.g., overall job satisfaction and job performance). For other situations where the constructs are expected to be related (e.g., overall job satisfaction and organizational commitment), the effect due to correlated transient error must be taken into account. As a result, the procedure described in this section cannot be applied.

The literature abounds with studies reporting correlations between measures for the same construct. However, most of the time, the correlations are obtained when the measures are administered on one occasion. As shown in Appendix A, such correlations account for the effects of random response error, item-specific factor error, and scale-specific factor error, but not transient error. We call this correlation the generalized coefficient of equivalence (GCE) because when compared with the GCES (which takes into account all measurement artifacts), it is analogous to the CE in classical test theory when the latter is compared with the CES.

Appendix A shows that

$$GCES = GCE - TEV, \tag{6}$$

where $\text{TEV} = Var(po)/Var(X)$ (proportion of transient error variance to observed variance).

The second component in the right side of Equation 6 above (TEV) can be estimated from the following equation provided by Schmidt et al. (2003):

$$TEV = CE - CES, \tag{7}$$

where CE = coefficient of equivalence (typically coefficient alpha or other coefficients for internal consistency); and CES = coefficient of equivalence and stability.

Equation 7 requires knowledge of the CE and CES of the measures of interest. However, CES is rarely estimated and reported in the literature. In the absence of a CES estimate, coefficient of stability (CS; Schmidt et al., 2003) can be used to compute a lower bound estimate of TEV:

$$TEV \geq TEV_{\min} = CE - CS, \tag{8}$$

where $\text{TEV}_{\min}$ is the lower bound of TEV. From Equations 6 and 8, we can derive an equation enabling estimation of the upper bound of GCES:

$$GCES \leq GCES_{\max} = GCE - TEV_{\min}, \tag{9}$$

where $\text{GCES}_{\max}$ is the upper bound of GCES.

Combining Equations 8 and 9, we have a formula to estimate the upper bound of the GCES for measures of a construct from the GCE, CE, and CS of those measures:

$$GCES_{\max} = GCE - (CE - CS). \quad (10)$$

As defined earlier, GCE is the correlation between different scales measuring the same underlying construct, CE is the coefficient alpha of the scales, and CS is the test-retest reliability. These reliability coefficients can be readily found in many studies involving the construct in the literature. They therefore can be meta-analytically estimated (e.g., Viswesvaran & Ones, 2000) and used to calculate an upper bound of the GCES of these measures.

Using the upper bound of the GCES to correct for measurement error is conservative because it still somewhat underestimates the true correlations between constructs. However, this provides better estimates than does the traditional approach of correcting for measurement error using the CE (coefficient alpha). The approach described here allows the accumulation over time of appropriate reliability estimates to be used in correcting for measurement artifacts using equation 2. Adding to this advantage, compared with the SEM approach, the GCES approach is also simpler computationally, so it is likely to be the method of choice for accurately estimating the true correlations among constructs. There is, in fact, precedent for this general cumulative approach. For example, Rothstein (1990) presented large-sample meta-analytically derived figures for interrater reliability of supervisory ratings of overall job performance, which have subsequently been used to make corrections for measurement error in ratings in many published studies.

In the next section, we present an application of the meta-analytic approach discussed here to estimate the GCES for measures of one important organizational construct: overall job satisfaction.

Estimating the GCES of Measures for Overall Job Satisfaction Construct

Transient Error in Measures of Job Satisfaction

The existence of transient error in measures of organizational constructs in general, and job satisfaction in particular, is certainly controversial. In the earlier section, we discussed theoretical reasons that transient error should be considered for job satisfaction measures. Yet, there remain two important empirical questions that need to be addressed before organizational researchers can fully embrace the concept of transient error. First, how can we know if the changes in people's responses to measures of job satisfaction across occasions reflect transient error or if they are due to real changes in the job satisfaction construct? Second, how large is the effect of transient error on job satisfaction measures? We attempt to address the first question in this section. The answer to the second question is provided in the next section, where the proportion of transient error variance to observed score variance (TEV) of job satisfaction measures is meta-analytically estimated.

Schmidt and Hunter (1996) discuss a simple approach that enables determining whether test-retest correlations reflect changes in the true constructs underlying a measure or just

transient error. The approach involves comparing three test-retest correlations for the measure across three occasions (r_{12}, r_{13}, and r_{23}). If these correlations are approximately equal (within allowance due to sampling error), it can be concluded that the construct underlying the measure is stable (unchanged) within the period from time 1 to time 3. This approach provides the important foundation for the search to answer the question. However, the simplicity of the approach limits information one can draw from it if differences in the test-retest correlations are detected. It is not clear if such differences are due to real changes in the construct or in other measurement artifacts.

The literature on measurement invariance (Vandenberg & Lance, 2000) suggests a solution that can augment the Schmidt and Hunter (1996) approach. By examining the invariance of factor structures of measurement models for observed scores across times, we can statistically determine if there are changes in the construct during the period. As suggested by Schmidt and Hunter, we need data for at least three occasions to conduct the analysis. We use data from a published study (Ilies & Judge, 2004) to demonstrate how such analysis can be done to address the question.

Sample. Ilies and Judge (2004) suggested an experience-sampling method to measure job satisfaction. Over a period of 2 weeks, the authors had participants who are employees of a state university reply to a five-item measure of overall job satisfaction (together with measures for other constructs) three times a day. The data are especially relevant for our purpose because (a) there are more than two occasions for the job satisfaction measure, and (b) the total period is relatively short (2 weeks), which can ensure that there was no big change in work environments that might lead to changes in the job satisfaction construct. We selected three occasions (day 1, day 9, and day 12) and averaged all three responses by each participant on each day to create indicators for job satisfaction for the participant on that day.[7] Our sample includes data from 44 participants, each providing responses to five job satisfaction items on three selected occasions.[8]

Methods. Based on the latent state-trait model (Schmitt & Steyer, 1993), we specified models with three factors (one for each occasion) representing the combined effect of transient error (state) and job satisfaction true score (trait) underlying participants' responses to the job satisfaction items (5 items $\times$ 3 occasions).[9] There is a common second-order factor representing the job satisfaction true score underlying these three factors. The disturbance terms for these first-order factors reflect the transient error. To capture the effect of item-specific factor error, we allow residuals for the same items across occasions to be correlated.[10] Following the measurement invariance approach (Vandenberg & Lance, 2000), we examined two hierarchically nested models. The first model allows all the parameters to be freely estimated. The second model constrained all the factor loadings (loadings of first-order factors on the items across occasions, and those of the second-order factor on the first-order factors) to be equal. It also constrained the three disturbances (reflecting transient error) to be equal. Comparing the second model with the first model (serving as the base model) allows examining invariance of the job satisfaction construct and transient error across the three occasions. Figure 2 shows the models. We used Proc CALIS (SAS 9.1) to analyze the data.

Results. Model 1 (base model) appears to marginally fit the data (CFI = .95, RSMEA = .10, $\chi^2 = 104.94$, $df = 72$, $p < .01$). Model 2 (constrained model) also fits the data equally (CFI = .95, RSMEA = .10, $\chi^2 = 130.17$, $df = 94$, $p < .01$). The disturbances [transient error variance *Var*(*po*)] are estimated to be .26, which is about 44% of the variance of the first-order factors, which represent the combined effects of job satisfaction true score and transient error. The chi-square difference comparing the two models is not statistically significant ($\Delta\chi^2 = 25.23$, $df = 22$, $p > .10$). Further comparing the models using the criterion of .01 for CFI difference as suggested by Cheung and Rensvold (2002) also yields the same results. Taken together, it appears that Model 2, which is more parsimonious, should be preferred and retained.

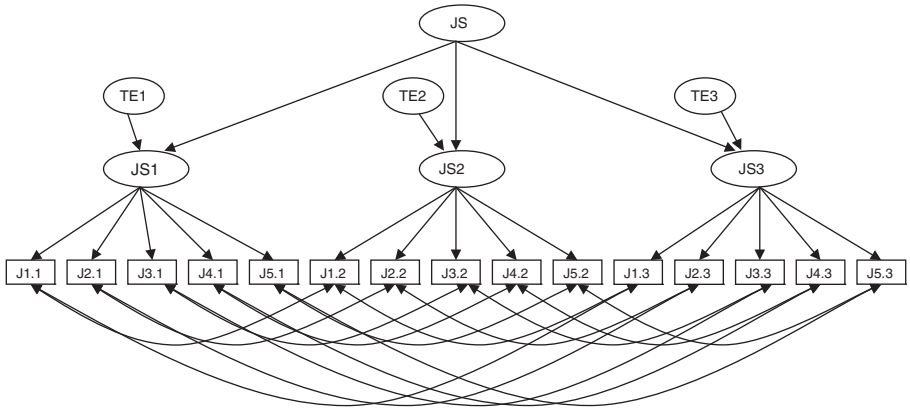
Discussion. Results of this analysis suggest that the construct of job satisfaction is stable during the period of approximately 2 weeks (12 days). It also shows that transient error contributes substantially to observed variance of the job satisfaction measure.[11] However, due to the small sample size used here, we do not claim that the results provide the conclusive answer to the question about the existence of transient error in measures of job satisfaction. Instead, as mentioned earlier, our purpose is just to present an approach that can help empirically examine the question. The result here should be treated as tentative, pending future replications.

Meta-Analytic Estimates of GCES for Job Satisfaction Measures

We conducted a literature search to identify studies that report (a) correlations among established measures of overall job satisfaction, (b) CE, and (c) CS (test-retest reliability) for those measures. Searches were conducted for the following scales: Job Descriptive Index (JDI; Smith, Kendall, & Hulin, 1969), JIG (Ironson et al., 1989), MSQ (Weiss et al., 1967), Brayfield and Rothe scale (Brayfield & Rothe, 1951), Hoppock scale (Hoppock, 1935), and Kunin's Faces scale (Kunin, 1955). An electronic search was carried out using the PsycINFO database with keywords being names of the selected scales. This search located 212 articles. A further search based on citations in the two comprehensive reviews of measures of organizational constructs (J. D. Cook, Hepworth, Wall, & Warr, 1981; Price, 1997) yielded 31 additional articles. For each article, reliability coefficients (and sample sizes) reported for the above scales (CS: test-retest and CE: alpha coefficient or split-half corrected by Spearman-Brown) were recorded (except for Kunin's Faces scale, which has no reports for reliability because there is only one item). Intercorrelations among the scales, whenever available, were recorded (here, data from Kunin's Faces scale were recorded).

Analysis. A series of meta-analyses (Hunter & Schmidt, 1990) was carried out to estimate the CE, CS, and GCE (intercorrelations of the measures) of the five job satisfaction measures (JDI, JIG, MSQ, Brayfield & Rothe, Hoppock) separately. For the JDI, CE was estimated for total scale (overall job satisfaction) as well as the five facet scales. Only CE for total scales was estimated for the other four measures. We were able to obtain enough reports of test-retest reliability to estimate CS only for JDI scales. Thus, only CS and CE

Figure 2
The Structural Equation Model for Estimating the Effect of Transient Error Underlying Responses to Job Satisfaction Measures



Note: Residuals for the indicators are not shown to simplify the presentation. J1.1 = responses to job satisfaction item 1 on occasion 1; J2.1 = responses to job satisfaction item 2 on occasion 1; J3.2 = responses to job satisfaction item 3 on occasion 2; J5.3 = responses to job satisfaction item 5 on occasion 3; TE1, TE2, TE3 = transient error on occasions 1, 2, and 3, respectively; JS1, JS2, JS3 = combined effect of job satisfaction and transient error on occasions 1, 2, and 3, respectively; JS = job satisfaction factor.

for JDI are used in Equation 8 to estimate the lower bounds of the transient error variance (TEV_{min}) of job satisfaction facets. The average of these values is then used to estimate the lower bound of transient error variance of overall job satisfaction. This value (TEV_{min}) is then used in Equation 9 together with the GCE (the average of cells in the intercorrelation matrix of the six selected scales) to estimate the upper bound of GCES of job satisfaction measures.

Results. Results of the meta-analyses for CE are shown in Table 5. There, it can be seen that CEs for all the scales are fairly high, ranging from .78 (for pay satisfaction subscale of JDI) to .92 (JIG) and averaging .87. Considering the dependence of the construct of job satisfaction on job environments that are not perfectly stable, we did two separate analyses for CS, one for short periods (less than 3 months) as defined by Harrison and Martocchio (1998) and another for all the available test-retest coefficients reported in the literature. Table 6 shows results for these analyses. As expected, mean CS computed for all time intervals is lower (ranging from .56 for supervisor satisfaction to .68 for overall satisfaction) than those including only short intervals (these range from .67 for promotion satisfaction to .77 for work satisfaction). Accordingly, we used only CS values obtained from test-retest coefficients with short intervals in estimating the lower bound of transient error variance (TEV_{min}). Results of this analysis are given in Table 7, which illustrates estimates of TEV_{min} for JDI job facets using Equation 8 with values of CE and CS obtained from Tables 5 and 6. These TEV_{min} values are then averaged to yield the estimate of the lower bound of transient error variance of overall job satisfaction. The result (see Table 7), .12, is rather high. Yet, the true TEV may be larger than this lower bound estimate.

Table 5
Coefficient of Equivalence (CE) for Popular Job Satisfaction Scales

Scale	*K* (# of cases)	*N* (total sample size)	CE Estimate	*SD* of CE
JDI—total (sum of subscales)	15	3,298	0.82	0.125
JDI—work	73	20,773	0.80	0.054
JDI—pay	48	14,711	0.78	0.046
JDI—supervisor	72	20,947	0.86	0.043
JDI—promotion	52	15,041	0.85	0.052
JDI—coworker	51	15,928	0.87	0.043
JIG	4	4,985	0.92	0.023
MSQ	51	8,965	0.89	0.030
Brayfield-Rothe	17	3,036	0.88	0.036
Hoppock	3	29,798	0.83	0.010
Average of overall job satisfaction (excluding JDI facets)			0.87	
Average of JDI facets			0.83	

Note: JDI = Job Descriptive Index (Smith, Kendall, & Hulin, 1969); JIG = Job In General (Ironson, Smith, Brannick, & Gibson, 1989); MSQ = Minnesota Satisfaction Questionnaire (Weiss, Dawis, England, & Lofquist, 1967); Brayfield-Rothe = job satisfaction scale (Brayfield & Rothe, 1951); Hoppock = job satisfaction scale (Hoppock, 1935).

Table 8 presents the intercorrelations among different job satisfaction measures. The average intercorrelation in the matrix is .70, and this is our estimate of the average GCE of these scales. From Equation 9, the upper bound of this coefficient can be obtained by subtracting TEV_{min} from GCE: $GCES_{max} = GCE - TEV_{min} = .70 - .12 = .58$.

It may be surprising to find that more than 42% of the observed variance of existing job satisfaction scales is due to measurement artifacts. Nevertheless, this value is only an approximate estimate of the GCES due to the limitations of the studies available for the meta-analyses (i.e., there were no reports of CES to enable our direct estimation of TEV, and the number of intercorrelations between different scales measuring the construct of job satisfaction was limited). More studies with appropriate designs directly addressing the topic are needed to strengthen the conclusion about the magnitude of the GCES coefficient.

Conclusion

The importance of the procedures presented in this article can be illustrated concretely by examining the following example. Let us assume that the coefficient alpha for a job satisfaction scale is .82 (cf. Peterson, 1994). The estimate of the more appropriate GCES for the scale is .58. Suppose the observed correlation between a job satisfaction scale and a measure of another construct is .40. Let us say, for the sake of illustration, that the coefficient alpha for the other scale is also .82 and its GCES is also .58. What are the respective estimates of the construct-level correlations? The researcher who uses the coefficient alpha in the disattenuation formula (Equation 2) to correct for measurement error obtains an estimate of .49; the researcher who uses the GCES obtains an estimate of .69, which is

Table 6
Test-Retest Reliability (Coefficient of Stability) for Job Descriptive Index (JDI) Subscales and Full Scale

	Short Intervals				All Intervals			
JDI	*r*	*k*	*N*	*SD*	*r*	*k*	*N*	*SD*
Work	0.77	5	364	0.08	0.61	15	2,268	0.11
Pay	0.71	5	364	0.07	0.65	9	1,425	0.06
Supervisor	0.71	5	364	0.08	0.56	11	1,703	0.14
Promotion	0.67	5	364	0.08	0.62	11	1,561	0.09
Coworker	0.69	5	364	0.05	0.59	10	1,649	0.12
Total					0.68	3	390	0.03

Note: Short interval = 21 to 63 days (average = 34 days); all intervals, including all reliabilities reported in the literature = 21-480 days (average = 155 days).

Table 7
Estimates of Lower Bound of Transient Error Variance (TEV_{min}) of Job Descriptive Index (JDI) Subscales

JDI	CE	CS	TEV_{min}
Work	0.80	0.77	0.03
Pay	0.78	0.71	0.06
Supervisor	0.86	0.71	0.15
Promotion	0.85	0.67	0.18
Coworker	0.87	0.69	0.18
Average	0.83	0.71	0.12

Note: Only test-retest reliability coefficients with a short interval are used for estimating TEV_{min}. CE = coefficient of equivalence; CS = coefficient of stability.

Table 8
Intercorrelations of Measures for Overall Job Satisfaction

	JIG				Brayfield				Hoppock				Kunin				MSQ			
	r	*N*	*k*	*SD*	*r*	*N*	*k*	*SD*	*r*	*N*	*k*	*SD*	*r*	*N*	*k*	*SD*	*r*	*N*	*k*	*SD*
Brayfield	.80	227	1																	
Hoppock					.86	271	2	.043												
Kunin	.75	227	1		.65	272	2	.000	.77	139	2	.026								
MSQ					.67	270	1						.71	144	1					
JDI	.66	227	1		.54	486	5	.090	.61	860	3	.042	.71	311	3	.042	.75	284	3	.024

Note: Average correlation: 0.70 (k = 12, SD = 0.09). JIG = Job In General (Ironson, Smith, Brannick, & Gibson, 1989); Brayfield-Rothe = job satisfaction scale (Brayfield & Rothe, 1951); Hoppock = job satisfaction scale (Hoppock, 1935); Kunin = Kunin's Faces scale (Kunin, 1955); MSQ = Minnesota Satisfaction Questionnaire (Weiss, Dawis, England, & Lofquist, 1967); JDI = Job Descriptive Index (Smith, Kendall, & Hulin, 1969); r = average of correlations between two measures of overall job satisfaction; N = total sample size of all the studies included to estimate the average of correlations; k = number of studies included to estimate the corresponding average of correlations; SD = standard deviation of correlations between two measures of job satisfaction.

29% larger. This is a substantial bias. The downward biases in the estimates produced by use of the inappropriate reliability coefficient are large enough to cause nontrivial errors in research conclusions.

To some, the corrections for measurement artifacts illustrated and demonstrated in this article may appear radical. One way to show that this is not the case is to again set our methods and proposals in the context of SEM methods. If in a SEM study, several different scales are used as indicators (measures) of a latent variable (construct) of interest, and each scale is measured on a different day, the correction for artifacts in the measurement of that construct would be identical to that described in this article. We have shown that the two procedures, GCES and SEM, are conceptually and computationally equivalent. Yet, most would not consider the SEM method unusual or worthy of comment; on the contrary, this practice would be considered exemplary. This makes clear that what we advocate in this article is consistent with the best current practice in SEM. However, most SEM applications fall short of this ideal. Even if different scales of a construct are used as indicators, they are typically all administered at the same time, thus failing to control for transient error. And often, the different ''scales'' are merely random subsets of items taken from the same longer scale, resulting in a correction similar to the use of the CE as the reliability index. But the point is that a SEM study corresponding to the method we advocate would rightly be viewed not only as acceptable but as exemplary.

The goal of scientific research is to understand relationships between constructs, not just between measures of these constructs. Correcting the observed correlations for downward biases created by measurement artifacts is required to achieve that goal. Doing this requires knowledge of measurement artifacts and appropriate choice of the reliability coefficient (Cronbach, 1947; Schmidt & Hunter, 1996, 1999; Schmidt et al., 2003; Thorndike, 1951). The nearly universal use of coefficient alphas or test-retest reliability falls short of correcting for all the sources of measurement artifacts inherent in the measures of psychological and organizational constructs. This article calls attention to the problem and provides the analytical tools to enable researchers to effectively deal with the biases created by the multifaceted nature of measurement artifacts. The ultimate goal is more accurate estimates of construct-level correlations, because such correlations are the building blocks of theory.

Appendix A
Variance Components Included in Different Types of Reliability Coefficients

The four major sources of measurement artifacts in self-report measures can be presented under generalizability theory in accordance with the three-facet partially nested random effects model: $p \times o \times (i : s)$.

Based on the model, a response X_{pois} of person p on occasion o on item i of scale s can be decomposed into the following components:

1. Overall mean $= \mu$,
2. Person effect $(p) = \mu_p - \mu$,
3. Occasion effect $(o) = \mu_o - \mu$,
4. Scale effect $(s) = \mu_s - \mu$,

5. Item within scale effect $(i:s) = \mu_{is} - \mu_s$,
6. Person $\times$ occasion effect $(po) = \mu_{po} - \mu_o - \mu_p + \mu$,
7. Person $\times$ scale effect $(ps) = \mu_{ps} - \mu_s - \mu_p + \mu$,
8. Occasion $\times$ scale effect $(os) = \mu_{os} - \mu_o - \mu_s + \mu$,
9. Person $\times$ item within scale effect $(pi:s) = \mu_{pis} - \mu_{is} - \mu_{ps} + \mu_s$,
10. Occasion $\times$ item within scale effect $(oi:s) = \mu_{ois} - \mu_{os} - \mu_{is} + \mu_s$,
11. Person $\times$ occasion $\times$ scale effect $(pos) = \mu_{pos} - \mu_{po} - \mu_{ps} - \mu_{os} + \mu_p + \mu_o + \mu_s - \mu$,
12. Residual effect $(poi:s, \varepsilon) = X_{pois} - \mu_{pos} - \mu_{pis} - \mu_{ois} + \mu_{ps} + \mu_{os} + \mu_{is} - \mu_s$,

where μ = overall mean: mean of all the responses across i, o, and s; μ_x = mean of responses on level x of facet X across all levels of the remaining facets (Y, Z, T); μ_{xy} = mean of responses on level x of facet X and level y of facet Y across all levels of the remaining facets (Z, T); μ_{xyz} = mean of responses on level x of facet X and level y of facet Y and level z of facet Z across all levels of the remaining facet (T). (x, y, z, t can be p, i, o, or s).
So,

$$\begin{aligned} X_{pois} &= (1) + (2) + (3) + (4) + (5) + (6) + (7) + (8) + (9) + (10) + (11) + (12) \\ &= \mu + p + o + s + i{:}s + po + ps + os + pi{:}s + oi{:}s + pos + (poi{:}s, \varepsilon). \end{aligned} \tag{A1}$$

We can group all the components that do not include any item with subscript p (i.e., they are averages across persons p, and thus independent of p) together and symbolize them C. Equation A1 can then be rewritten as

$$X_{pois} = p + po + ps + pi{:}s + pos + (poi{:}s, \varepsilon) + C. \tag{A2}$$

Furthermore, by grouping pos and $(poi{:}s, \varepsilon)$ and symbolizing this component e, we have

$$X_{pois} = p + po + ps + pi{:}s + e + C. \tag{A3}$$

As such, the observed score X of person p includes six components: (a) construct score (p), (b) interaction between persons and occasions (po = transient error), (c) interaction between persons and scales (ps = scale-specific factor error), (d) interaction between persons and items within scales ($pi{:}s$ = item-specific factor error), (e) residual (e = random response error), and (f) C, whose value is a constant for all persons p.

Given the usual assumption of independence, the observed variance of a measure (across persons) is therefore

$$Var(X) = Var(p) + Var(po) + Var(ps) + Var(pi:s) + Var(e). \tag{A4}$$

The Generalized Coefficient of Equivalence and Stability

As mentioned in the text, the generalized coefficient of equivalence and stability (GCES) is the appropriate coefficient to be used in the disattenuation formula, so it should have the form of

$$GCES = Var(p)/Var(X) \tag{A5}$$

(i.e., there should not be any components of measurement artifacts in the numerator of the formula for GCES). GCES can be estimated by correlating two scales (α and β) of the same construct

administered on two different occasions (1 and 2). Because scales α and β are congeneric (Jöreskog, 1971), responses to scales α at time 1 and β at time 2 can generally be written as

$$X_\alpha = k_\alpha p + h_\alpha po_1 + l_\alpha ps_\alpha + m_\alpha pi : s_\alpha + e_{1\alpha} + C_{1\alpha}, \tag{A6}$$

and

$$X_\beta = k_\beta p + h_\beta po_2 + l_\beta ps_\beta + m_\beta pi : s_\beta + e_{2\beta} + C_{2\beta}, \tag{A7}$$

where k_α, h_α, l_α, m_α, and C_α are attributes of scale α; k_β, h_β, l_β, m_β, and C_β are attributes of scale β.

To simplify our presentation, we combine all the measurement error components on the right sides of Equations A6 and A7 and symbolize them E_α and E_β, respectively. Equations A6 and A7 then become

$$X_\alpha = k_\alpha p + E_\alpha, \tag{A8}$$

and

$$X_\beta = k_\beta p + E_\beta. \tag{A9}$$

From Equations A5, A8, and A9, the GCES of scales α and β can be written as

$$GCES_\alpha = Var(p_\alpha)/Var(X_\alpha) = Var(k_\alpha p)/Var(X_\alpha) = k_\alpha^2 Var(p)/Var(X_\alpha). \tag{A10}$$

and

$$GCES_\beta = Var(p_\beta)/Var(X_\beta) = Var(k_\beta p)/Var(X_\beta) = k_\beta^2 Var(p)/Var(X_\beta). \tag{A11}$$

Also, from Equations A8 and A9, the correlation between the scales can be written as

$$R_{\alpha\beta} = \frac{Cov(k_\alpha p + E_\alpha, k_\beta p + E_\beta)}{\sqrt{Var(X_\alpha)Var(X_\beta)}}. \tag{A12}$$

From the assumption that measurement errors are independent of one another and of true score, the numerator on the right side of Equation A12 can be expanded and simplified. Equation A12 then becomes

$$R_{\alpha\beta} = \frac{k_\alpha k_\beta Var(p)}{\sqrt{Var(X_\alpha)Var(X_\beta)}}. \tag{A13}$$

From Equations A10 and A11, Equation A13 can be rewritten as

$$R_{\alpha\beta} = \sqrt{GCES_\alpha GCES_\beta}. \tag{A14}$$

Equation A14 shows that the correlation between scales α and β (when they are administered on two different occasions) is the geometric mean of the GCESs of these scales.
If we assume that $GCES_\alpha = GCES_\beta$, Equation A14 will become

$$R_{\alpha\beta} = GCES_\alpha = GCES_\beta. \tag{A15}$$

Thus, the GCES of scales α and β can be estimated by their correlation $R_{\alpha\beta}$.

The Generalized Coefficient of Equivalence (GCE)

From Equations A6 and A7, we can write the correlation between scales α and β (when they are administered on the same occasion, after simplifying components in the numerator that are uncorrelated by definition) as

$$R_{\alpha\beta} = GCE = \frac{k_\alpha k_\beta Var(p) + h_\alpha h_\beta Var(po)}{\sqrt{Var(X_\alpha)Var(X_\beta)}}. \tag{A16}$$

From Equations A13 and A14, the assumption that $GCES_\alpha = GCES_\beta$, stated above, can be rewritten as

$$k_\alpha^2 / Var(X_\alpha) = k_\beta^2 / Var(X_\beta) \tag{A17}$$

Solving Equation A17 for $Var(X_\beta)$ and inserting it into Equation A16, we have

$$GCE = \frac{k_\alpha^2 Var(p) + \frac{k_\alpha h_\alpha h_\beta}{k_\beta} Var(po)}{Var(X_\alpha)}. \tag{A18}$$

We further assume that the proportion of variance of transient error in the measures is proportional to the proportion of their true score variance $Var(p)$, that is,

$$\frac{Var(k_\alpha p)}{Var(k_\beta p)} = \frac{Var(h_\alpha po)}{Var(h_\beta po)} \text{ or } \frac{k_\alpha^2 Var(p)}{k_\beta^2 Var(p)} = \frac{h_\alpha^2 Var(po)}{h_\beta^2 Var(po)} \text{ or } \frac{k_\alpha}{k_\beta} = \frac{h_\alpha}{h_\beta} \tag{A19}$$

Solving Equation A19 for k_β and inserting it into Equation A18, we have

$$GCE = \frac{k_\alpha^2 Var(p) + h_\alpha^2 Var(po)}{Var(X_\alpha)} = \frac{Var(k_\alpha p) + Var(h_\alpha po)}{Var(X_\alpha)} = \frac{Var(p_\alpha) + Var(po_\alpha)}{Var(X_\alpha)}. \tag{A20}$$

Following similar procedures described above, we can also show that

$$GCE = \frac{Var(k_\beta p) + Var(h_\beta po)}{Var(X_\beta)} = \frac{Var(p_\beta) + Var(po_\beta)}{Var(X_\beta)}. \tag{A21}$$

Equations A20 and A21 show that GCE, which is the correlation between two measures of a construct administered on the same occasion, contains in its numerator not only the true score variance but also the transient error variance.

Appendix B
Examining the Bias of the Generalized Coefficient of Equivalence and Stability (GCES) Procedure Due to Violation of Its Assumption of Equivalence of Measures for the Same Construct

Assume we have two measures, scale A and scale B, for construct K. The GCESs of these measures are $GCES_A$ and $GCES_B$, respectively. Call ρ_{KL} construct-level correlation between construct

K and construct L. The observed correlations between scale A and scale B with a measure of construct L are

$$r_A = \rho_{KL}\sqrt{GCES_A GCES_L}, \tag{B1}$$

and

$$r_B = \rho_{KL}\sqrt{GCES_B GCES_L}; \tag{B2}$$

where r_A = observed correlations between scale A and the measure of construct L; r_B = observed correlations between scale B and the measure of construct L; and $GCES_L$ = the GCES of the measure of construct L.
Without loss of generality, we assume $GCES_L = 1$. Then, Equations B1 and B2 can be rewritten as follows:

$$r_A = \rho_{KL}\sqrt{GCES_A}, \tag{B3}$$

and

$$r_B = \rho_{KL}\sqrt{GCES_B}. \tag{B4}$$

Our estimate for the GCES of scales A and B ($\overline{GCES}_K$) is their correlation. As shown in Appendix A (Equation A14), this correlation is equal to the geometric mean of $GCES_A$ and $GCES_B$:

$$\overline{GCES}_K = r_{AB} = \sqrt{GCES_A GCES_B}, \tag{B5}$$

where r_{AB} = correlation between scale A and scale B when they are administered on two different occasions.

When applying the GCES method, we correct the mean observed correlation between the measures of construct K and the measure of construct L:

$$\hat{\rho}_{KL} = \frac{\bar{r}}{\sqrt{\overline{GCES}_K GCES_L}} = \frac{\bar{r}}{\sqrt{\overline{GCES}_K}}, \tag{B6}$$

where $\hat{\rho}_{KL}$ = estimated construct-level correlation between K and L by the GCES approach; and

$$\bar{r} = \frac{r_A + r_B}{2}.$$

From Equations B3, B4, and B5, Equation B6 can be rewritten as follows:

$$\hat{\rho}_{KL} = \frac{\frac{\rho_{KL}\sqrt{GCES_A} + \rho_{KL}\sqrt{GCES_B}}{2}}{\sqrt[4]{GCES_A GCES_B}} = \rho_{KL}\left(\frac{\sqrt{GCES_A} + \sqrt{GCES_B}}{2\sqrt[4]{GCES_A GCES_B}}\right). \tag{B7}$$

From Equation B7, we can calculate the percentage bias for the GCES method when its assumption of equivalence of measures is violated:

$$\%\text{Bias} = 100\frac{\hat{\rho}_{KL} - \rho_{KL}}{\rho_{KL}} = 100\left(\frac{\rho_{KL}\left(\frac{\sqrt{GCES_A} + \sqrt{GCES_B}}{2\sqrt[4]{GCES_A GCES_B}}\right) - \rho_{KL}}{\rho_{KL}}\right). \tag{B8}$$

Simplifying Equation B8, we have

$$\%\text{Bias} = 100\frac{\left(\sqrt[4]{GCES_A} - \sqrt[4]{GCES_B}\right)^2}{2\sqrt[4]{GCES_A GCES_B}}. \tag{B9}$$

Equation B9 shows that the GCES method yields a positively biased estimate of the construct-level correlation when its assumption of factor equivalence is violated (that is, when $GCES_A \neq GCES_B$).

From Equation B7, we can infer the more general situation where the construct-level correlation between K and L is estimated by scales A and B of construct K, and scales C and D of construct L:

$$\hat{\rho}_{KL} = \rho_{KL}\left(\frac{\sqrt{GCES_A} + \sqrt{GCES_B}}{2\sqrt[4]{GCES_A GCES_B}}\right)\left(\frac{\sqrt{GCES_C + GCES_D}}{2\sqrt[4]{GCES_C GCES_D}}\right). \tag{B10}$$

Notes

1. As noted earlier, our conceptualization of measurement artifacts is based on generalizability theory (Cronbach et al., 1972). Because one of the tenets of generalizability theory concerning the independence of variance components from different measurement facets can be seen as an extension from classical test theory (Feldt & Brennan, 1989), we also make extensive references to classical test theory to facilitate our presentations of the concepts.

2. This is the equation of observed variance across persons for a measure administered on one occasion. As such, it does not include variances due to scales [*Var*(s)] and occasions [*Var*(o)]. It should be noted that measurement artifacts are the interactions between persons and the sources of artifacts (occasions, scales, or items), so they always exist in observed variance, even though the measure in question is administered on one occasion, by one scale, or even with a single item (see Appendix A for details). Of course, under such conditions, the magnitude of these variance components cannot be estimated.

3. By this definition, transient errors for a person are independent across occasions. On the same occasions, however, transient errors for measures of different but related constructs may be correlated (e.g., bad mood can similarly influence responses to measures of job satisfaction and organizational commitment). This potential correlation violates the assumption of classical test theory and generalizability theory. We discuss the implication of this potential violation and solution in the "Assumptions Underlying the GCES Correction Procedure" section.

4. We generated data so that the correlations between transient errors of two measures are the same on both occasions because it is theoretically reasonable and consistent with the multivariate generalizability theory model (Brennan, 2001). However, it is not necessary to assume this invariance because use of cross-occasion correlations in the disattenuation formula as suggested here renders such differences in correlations irrelevant. To confirm this, we simulated data with different correlations between transient errors for different occasions (not reported here). Results confirm that the correction procedures discussed here (both GCES and SEM discussed later) do not require the correlations to be the same.

5. This is the $p \times (i : s/o)$ design in generalizability theory, with persons being crossed with items nested within scales that are totally confounded with occasions.

6. The simulation and estimation programs are available from the authors upon request.

7. We selected these days to maximize our sample size. Furthermore, the intervals between the days are reasonable to expect that transient errors are not correlated across occasions.

8. Our sample is slightly larger than that used by Ilies and Judge (2004) because the original data include missing responses on many occasions within the period due to attrition. As mentioned in note 7, we selected the three occasions that maximized our sample size.

9. It should be noted that this trait factor still includes scale-specific factor error because items from one scale are included in the model. However, this does not affect our examination of the effect of transient error.

10. The original latent state-trait model specifies latent factors to represent the effects. That approach is theoretically advantageous to the current approach because it unambiguously identifies the factors that account for the correlated errors (Gerbing & Anderson, 1984). However, here we are interested in true score and transient error, so specifying correlated residuals helps simplify the models and facilitate presentation. For models with three correlations underlying each factor (as in this case), results (estimates for other parameters and fit indexes) are exactly the same for both approaches (i.e., specifying correlated errors or latent factors).

11. The model examined here can be used to directly estimate the proportion of transient error variance to observed score variance (TEV). However, we did not attempt to estimate the TEV here because of the small sample size.

References

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait-multimethod matrices: The case of self-reported affects and perceptions at work. *Journal of Applied Psychology*, *75*, 547-560.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5*, 370-379.

Becker, T. E., & Cote, J. A. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management*, *20*, 625-641.

Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology*, *35*, 307-311.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brooke, P. P., Russell, D. W., & Price, J. L. (1988). Discriminant validation of measures of job satisfaction, job involvement, and organizational commitment. *Journal of Applied Psychology*, *73*, 139-145.

Campbell, D. T., & Fiske, M. R. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin*, *56*, 818-850.

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology*, *67*, 691-700.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement*, *14*, 183-196.

Conway, J. M. (1998). Estimation and uses of the proportion of method variance for multitrait-multimethod data. *Organizational Research Methods*, *1*, 209-222.

Cook, T. D. (1985). Postpositivist critical multiplism. In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21-62). Newbury Park, CA: Sage.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally.

Cook, J. D., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The experience of work: A compendium and review of 249 measures and their use*. London: Academic Press.

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika*, *12*, 1-16.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, *3*, 412-423.

Doty, H. D., & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research Methods*, *4*, 374-406.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Fuller, W. A. (1987).*Measurement error models*. New York: John Wiley.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*, *11*, 572-580.

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, *64*, 1029-1041.

Harris, M. M., & Bladen, A. (1994). Wording effects in the measurement of role conflict and role ambiguity: A multitrait-multimethod analysis. *Journal of Management*, *20*, 887-901.

Harrison, D. A., & Martocchio, J. J. (1998). Time for absenteeism: A 20-year review of origins, offshoots, and outcomes. *Journal of Management*, *24*, 305-350.

Hom, P. W., Caranikas-Walker, F., Prussia, G. E., & Griffeth, R. W. (1992). A meta-analytical structural equations analysis of a model of employee turnover. *Journal of Applied Psychology*, *77*, 890-909.

Hom, P. W., & Griffeth, R. W. (1991). Structural equations modeling test of a turnover theory: Cross-sectional and longitudinal analyses. *Journal of Applied Psychology*, *76*, 350-366.

Hoppock, R. (1935). *Job satisfaction*. New York: Harper Brothers.

Houts, A. C., Cook, T. D., & Shadish, W. R. (1986). The person-situation debate: A critical multiplist perspective. *Journal of Personality*, *54*, 52-105.

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

Ilies, R., & Judge, T. A. (2004). An experience-sampling measure of job satisfaction and its relationships with affectivity, mood at work, job beliefs, and general job satisfaction. *European Journal of Work and Organizational Psychology*, *13*, 367-389.

Ironson, G. H., Smith, P., Brannick, M. T., & Gibson, W. M. (1989). Construction of a Job In General scale: A comparison of global, composite, & specific measures.*Journal of Applied Psychology*, *74*, 193-200.

James, L. R., Demaree, R. G., Mulaik, S. A., & Ladd, R. T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology*, *77*, 3-14.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109-133.

Kenny, D. A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 111-124). Hillsdale, NJ: Lawrence Erlbaum.

Kunin, T. (1955). The construction of a new type of job satisfaction measure. *Personnel Psychology*, *8*, 65-77.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods*, *7*, 228-244.

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

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions.*Applied Psychological Measurement*, *13*, 335-361.

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling*, *1*, 116-146.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*, *73*, 107-117.

Mathieu, J. E., & Farr, J. L. (1991). Further evidence for the discriminant validity of measures of organizational commitment, job involvement, and job satisfaction. *Journal of Applied Psychology*, *76*, 127-133.

Meyer, J. P., Allen, N. J., & Smith, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component conceptualization. *Journal of Applied Psychology*, *78*, 538-551.

Mowday, R., Steers, R., & Porter, L. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, *14*, 224-247.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*, 381-391.

Price, J. L. (1997). Handbook of organizational measurement. *International Journal of Manpower*, *18*, 303-558.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods*, *9*, 99-112.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, *75*, 322-327.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

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

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual difference constructs. *Psychological Methods*, *8*, 206-224.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personality Psychology*, *53*, 901-912.

Schmitt, M. J., & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, *14*, 519-529.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 356-442). Washington, DC: American Council on Education.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: John Wiley.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69.

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

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis: University of Minnesota, Industrial Relations Center.

Williams, L. J. (1995). Covariance structure modeling in organizational research: Problems with the method versus applications of the method. *Journal of Organizational Behavior*, *16*, 225-233.

Williams, L. J., Cote, J. A., & Buckley, M. R. (1989). Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology*, *74*, 462-468.

Huy Le is an assistant professor in the Department of Psychology at the University of Central Florida. His research interests include personnel selection, cross-cultural issues, psychometrics, and quantitative research methods (meta-analysis, Monte Carlo simulation). He has published in journals such as *Psychological Bulletin*, *Psychological Methods*, *Journal of Applied Psychology*, *Personnel Psychology*, *Personality and Social Psychology Bulletin*, and *Educational and Psychological Measurement*. He received his PhD in human resource management at the University of Iowa in 2003.

Frank L. Schmidt has been the Ralph L. Sheets Professor of Human Resources in the Tippie College of Business at the University of Iowa since 1985. He received his doctorate in industrial/organizational psychology from Purdue University in 1970. He has been an assistant and associate professor of I/O psychology at Michigan State, and for 11 years (1974–1985), he directed a research program in employment selection at the U.S. Office of Personnel Management in Washington, D.C., where he was one of the two co-inventors of validity generalization methods. His interest and research areas include personnel testing, selection, and placement; the role of intelligence and personality in job performance; causal models of job performance; and research methodology, in particular meta-analysis methods and measurement issues. He is the coauthor of a widely cited text on meta-analysis. He has published more than 150 journal articles and book chapters in these areas and has received the Distinguished Scientific Contributions Award (with John Hunter) from the American Psychological Association (APA) and (separately) from the Society for Industrial and Organizational Psychology (SIOP). He also received the Distinguished Career Award from the Human Resources Division of the Academy of Management, the Distinguished Career Achievement Award for Contributions to Research Methods from the Research Methods Division of the Academy of Management, and the Michael R. Losey Human Resources Research Award from the Society for Human Resource Management (SHRM). He is a fellow of

the APA, APS, and SIOP and is a past-president of the Measurement, Statistics, and Evaluation Division of the APA. He has been a consultant to many organizations on issues related to personnel selection.

Dan J. Putka is a senior scientist at the Human Resources Research Organization. He received his PhD and MS in industrial and organizational psychology from Ohio University. His research interests include personnel selection, modeling person-environment fit, employee turnover, and applied measurement issues.